# Studies of piRNA Synthesis

Muhan Wang

Department of Biochemistry

and

Lincoln College,

University of Oxford




A thesis submitted in partial fulfilment of the requirements for degree of Doctor of Philosophy at the University of Oxford

Michaelmas 2011

# Abstract

## Studies of piRNA Synthesis


## Muhan Wang

Submitted for the degree of Doctor of Philosophy
Department of Biochemistry and Lincoln College,
University of Oxford
Michaelmas 2011



RNA silencing is a form of post-transcriptional gene regulation, in which a small RNA guides a member of the Argonaute protein family in an effector complex to repress target gene expression. piRNAs, found in germ cells, are the most recently discovered major subset of small RNAs. A key known function of piRNA is to repress the transposable elements in the germline and maintain the germline genome integrity. The defining features of the piRNAs are 1) they are ubiquitously methylated at the 3'-end of the 2'-OH group by methyltransferase Hen1; 2) they associate exclusively with the Piwi subfamily Argonaute proteins. Much is not understood about the biogenesis and the regulation of the piRNA pathway. One of the fundamental questions is how the 3'-end of the piRNA is generated and recognised specifically by Piwi but not by Ago subfamily Argonaute proteins. In this thesis, the high resolution crystal structure of the Aubergine PAZ domain, a domain from a Piwi subfamily Argonaute, bound to a 7 mer single-stranded methylated piRNA 'mimic', reveals the mode of recognition for the 3'-end of piRNAs by Piwi subfamily Argonautes. The structure provides the molecular basis for why Piwi but not Ago PAZ domains preferentially bind to RNAs with 2'-O-methylation at the 3'-end, thus conferring substrate specificity. The structural results are confirmed by biochemical studies. Biochemical and biophysical studies on the methyltransferase Hen1 have provided insights into substrate specificity for piRNA 3'-end methylation and revealed a potential regulatory role for the C-terminal region of the protein. Extensive biochemical analysis defined a minimal active Hen1/short RNA complex, though crystallisation screening yielded no crystals for structure determination. Overall, this study provides insights into the generation and molecular recognition of the piRNA in animals.

# Acknowlegements

I would like to thank first and foremost my supervisor James Parker for not only his continuous and thorough supervisions and support throughout my DPhil, but also introducing and guiding me to an exciting field of scientific research. I would also like to thank my second supervisor Martin Nobel for helping me with the technique of crystallography and Jane Endicott for first introducing me to structural biochemistry.

Throughout the duration of my DPhil, many people have helped me with various aspects of my research and to whom I am extremely grateful to: Keiko Yata first introduced me to the LMB and taught me various lab techniques; Sonja Lorenz, despite busy preparing for *viva*, has taken time to not only help me out in the lab many times at all sorts hours, but also showed me the life of Oxford; I must thank Alison Hole, Camilla Oxley, Sushmita Borthakur, Michel Erat, Ioannis Vakonakis, Elizabeth Anscombe, Cynthia Tsao and many others for making lab work especially at the evenings and weekends more exciting, and always happy to give me advice. In addition, I also thank Ed Lowe for first introducing me to data collection and processing, Paul Emsley for showing me all kinds of functions (I would otherwise never know) in Coot. Overall, I am grateful to all members of the LMB for creating a great working environment for the past years.

During my studies in Oxford, I cannot thank my friends enough for making my time here truly special, for supporting me throughout: Sam, Lee (Stan), Alistair, Wan, Chris, Bindy, Sarah, Clare, Grainne, James… Though not in Oxford, I am completely grateful to the C201 team, especially Andrew and

Andy, whom have helped to make my first job so enjoyable that it was possible for me to juggle a full time job and complete thesis writing at the same time.

To James, words are not enough to express my gratitude, for your tremendous help with my thesis writing, for giving me love and support, for being in my life through the ups and downs and for taking care of me whenever needed. 谢谢你。

Last but not least, I dedicate this thesis to my family, whom I owe everything to and love unconditionally. 妈妈，爸爸， 奶奶， 谢谢您们多年来对我不懈的支持与鼓励，给予我无限的爱护与关怀， 给我创造了最好的条件，无论是在生活和事业上，给我了无数的指点。妈妈，您是我尤其需要感激的人，无论是学术还是生活各个方面，你对我的影响和帮助都是无法衡量的。我最亲爱的家人，我的任何造就都归于你们。

# Abbreviations

| | |
|---|---|
| C. *elegans* | Caenorhabditis elegans |
| C-Terminal | Carboxyl Terminal |
| CCP4 | Collaborative Computational Project number 4 |
| COOT | Crystallographic Object-Oriented Toolkit |
| DGCR8 | DiGeorge Syndrome critical regions gene 8 |
| DNA | Deoxyribonucleic acid |
| dNTP | Deoxyribonucleotide triphosphate |
| dsRBD | double-stranded RNA binding domain |
| dsRNA | double-stranded RNA |
| DTT | Dithiothreitol |
| E. coli | Escherichia coli |
| EMSA | Electrophoretic Mobility Shift Assay |
| ESRF | European synchrotron radiation facility |
| GST | Glutathione S-Transferease |
| hcsiRNA | heterochromatic siRNAs |
| HEN1 | HUA Enhancer 1 |
| HEPES | N–[2 hydorxoethyl] piperazine-N-[2-ethansulponic acid] |
| IPTG | Isopropyl-β-s-thiogalactopyranoside |
| ITC | Isothermal titration calorimetry |
| kDa | kilo Dalton |
| LB | Luria Bertaini Broth |
| M | Molar |

| | |
|---|---|
| mRNA | Messenger RNA |
| miRNA | Micro RNA |
| MID | The middle domain |
| N-Terminal | Amino terminal |
| nat-siRNA | natural antisense transcript-derived siRNA |
| Nt | Nucleotide |
| PAGE | Polyacrylamide gel electrophoresis |
| PAZ | Piwi, Argonaute, Zwille |
| PCR | Polymer chain reaction |
| PDB | Protein Data Bank |
| piRNA | Piwi- interacting RNA |
| Piwi | P-element induced wimpy testis |
| PNK | Polynucleotide Kinase |
| PRMT5 | Protein Arginine N-methyltransferase 5 |
| rasiRNA | repeat associated small interfering RNA |
| RISC | RNA Induced Silencing Complex |
| RLC | RISC Loading Complex |
| rmsd | Root Mean Squared Distance |
| RNA | Ribonucleic acid |
| RNAi | RNA interference |
| RPM | Revolutions per minute |
| S. *pombe* | Schizosaccharomyces pombe |
| SAH | S-adenosyl-l-Homocysteine |
| SAM | S-adenosyl-l-Methionine |
| SDS | Sodium dodecly sulphate |

| | |
|---|---|
| siRNA | Small interfering RNA |
| ssRNA | single stranded RNA |
| ta-siRNA | trans-acting siRNA |
| TRBP | TAR RNA binding protein |
| UTR | Untranslated region |

Table of Contents

**Table of Contents** ..... viii

**List of Figures** ..... xii

**List of Tables** ..... xv

**1 Introduction** ..... 1
**1.1 Overview of RNA silencing** ..... 1
1.1.1 Small RNAs as regular molecules ..... 1
1.1.2 Argonaute proteins – the common theme ..... 2
1.1.3 Overview of Argonaute structure and mechanism ..... 5
1.1.4 Three major classes of small RNA silencing pathway ..... 7
1.1.4.1 miRNAs ..... 8
1.1.4.2 siRNAs ..... 10
1.1.4.3 piRNAs ..... 13
**1.2 piRNAs** ..... 14
1.2.1 The discovery of piRNAs ..... 14
1.2.2 Functions of piRNAs ..... 16
1.2.3 piRNA biogenesis ..... 19
1.2.3.1 Primary processing pathway ..... 20
1.2.3.2 Ping-Pong amplification cycle ..... 22
1.2.4 piRNA regulation by Tudor containing proteins ..... 24
1.2.5 3' methylation of piRNAs (and *Drosophila* siRNAs) by Hen1 proteins ..... 25
**1.3 The focus and the aim of the thesis: piRNA methylation** ..... 29

**2 Methods and Materials** ..... 31
**2.1 Introduction** ..... 31
**2.2 Specific Materials and methods for Hen1 and Aubergine PAZ** ..... 31
2.2.1 Cloning ..... 31
2.2.1.1 Hen1 ..... 31
2.2.1.2 Cloning of PAZ constructs ..... 33
2.2.1.3 Site – Directed Mutagenesis ..... 33
2.2.2 Protein expression and purification ..... 34
2.2.2.1 General ..... 34
2.2.2.2 Hen1 proteins ..... 35
2.2.2.3 PAZ ..... 36
2.2.3 Biochemical and biophysical assays ..... 37
2.2.3.1 Hen1 in *vitro* methylation assays ..... 37
2.2.3.1.1 With detection by autoradiography ..... 37
2.2.3.1.2 With detection by liquid scintillation counting ..... 38
2.2.3.2 Analytical gel filtration ..... 38
2.2.3.3 Analytical Ultracentrifugation ..... 39
2.2.3.4 Isothermal titration calorimetry (ITC) ..... 39
2.2.3.5 Electrophoretic mobility shift assay (EMSA) ..... 39
2.2.3.6 Fluorescence Anisotropy ..... 40
2.2.4 RNA oligonucleotides ..... 40

2.2.4.1 Hen1 ..... 40
2.2.4.2 PAZ ..... 41
2.2.5 Crystallization trials ..... 41
2.2.5.1 96-well robotic screens ..... 41
2.2.5.2 24-well screens ..... 42
**2.3 General materials and Methods ..... 42**
2.3.1 Transformation ..... 42
2.3.2 Polymerase Chain Reaction - PCR ..... 43
2.3.3 Restriction enzyme digest and ligation ..... 43
2.3.4 Western Blotting ..... 44
2.3.5 SDS-PAGE ..... 45
2.3.6 Websites for disorder prediction and secondary structure prediction ..... 45
2.3.7 Protein and nucleic acid concentration determination ..... 46
2.3.7.1 Protein concentration determination ..... 46
2.3.8 Nucleic acid concentration determination ..... 47
2.3.9 Small scale expression and solubility testing ..... 47
2.3.10 3C-Protease and Sumo Protease ..... 48
2.3.11 Limited proteolysis ..... 48
2.3.12 N-terminal sequencing ..... 48
2.3.13 Mass spectrometry ..... 49

**3 Results and Discussion ..... 50**
**3.1 Hen1 Expression, Purification and Crystallisation Trials ..... 50**
3.1.1 Introduction ..... 50
3.1.2 Strategy ..... 52
3.1.3 Initial expression studies ..... 52
3.1.3.1 Choice of species ..... 52
3.1.3.2 Purification of full length dHen1 ..... 53
3.1.4 Truncated constructs for dHen1 ..... 54
3.1.4.1 RONN disorder prediction ..... 54
3.1.4.2 Secondary structure prediction and definition of constructs boundaries ..... 55
3.1.4.3 Expression and solubility testing ..... 56
3.1.5 Construct of dHen1.4$_{(1-356)}$ (ΔC1) ..... 57
3.1.5.1 Expression ..... 57
3.1.5.2 Purification ..... 59
3.1.5.3 Limited Proteolysis ..... 62
3.1.5.4 N-terminal sequencing ..... 63
3.1.5.5 Crystallization trials ..... 64
3.1.6 Revisiting full length dHen1 ..... 64
3.1.6.1 Purification of full length dHen1 construct dHen1.2$_{(1-391)}$ ..... 64
3.1.6.2 Limited Proteolysis ..... 64
3.1.6.3 N-terminal sequencing of limited proteolysis product ..... 65
3.1.6.4 Determination of the C-terminal truncation position ..... 66
3.1.7 Testing constructs ΔC5 - dHen1.12$_{(1-368)}$ and ΔC6 - dHen1.14$_{(1-364)}$ ..... 67
3.1.7.1 Cloning, expression and purification of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ ..... 67
3.1.7.2 Comparison with proteolysis ..... 69
3.1.8 Crystallization trials of the new constructs ..... 71
3.1.9 Discussion ..... 71
**3.2 Hen1 Mechanisms and Activity ..... 72**
3.2.1 Introduction ..... 72
3.2.2 Methylation of short RNAs in vitro ..... 75

3.2.2.1 Analysis by autoradiography ..... 75
3.2.2.2 Analysis by scintillation counting ..... 81
3.2.3 Interaction with Argonaute ..... 84
3.2.4 Impact of C-terminal tail on confirmation ..... 85
3.2.5 Discussion ..... 87
**3.3 PAZ – methyl piRNA protein structure: cloning, purification, crystallisation and structure determination ..... 90**
3.3.1 Introduction ..... 90
3.3.2 Construct design ..... 93
3.3.3 Expression and solubility testing ..... 94
3.3.4 Purification of AubPAZ3$_{(279-398)}$ ..... 95
3.3.5 Crystallisation of AubPAZ3$_{(279-398)}$ ..... 96
3.3.5.1 AubPAZ3$_{(279-398)}$/ RNA1 complex formation ..... 97
3.3.5.2 Initial Screening ..... 97
3.3.5.3 Optimisation ..... 98
3.3.6 Crystal harvesting, data collection and crystal structure determination ..... 99
3.3.6.1 Crystal harvesting ..... 99
3.3.6.2 Data collection ..... 99
3.3.6.3 SelenoMet AubPAZ3$_{(279-398)}$ expression purification and crystallisation ..... 100
3.3.6.4 Structure Determination by Molecular Replacement ..... 100
3.3.6.5 Structure Solutions ..... 100
3.3.7 Manual model building and refinement ..... 101
3.3.8 Discussion ..... 104
**3.4 Structure Analysis of AubPAZ3$_{(279-398)}$ / piRNA complex ..... 105**
3.4.1 Introduction ..... 105
3.4.2 Aubergine Argonaute PAZ domain – RNA structure ..... 106
3.4.2.1 Overall structure of AubPAZ3$_{(279-398)}$ with single stranded 2'-O-methyl RNA ..... 106
3.4.2.2 Two molecules in the asymmetric unit ..... 110
3.4.2.3 Detailed recognition of RNA binding – involvement of conserved residues ..... 111
3.4.2.4 Recognition of 2'-O-methyl group in a hydrophobic pocket ..... 115
3.4.2.5 A 'flipped' nucleotide ..... 116
3.4.2.6 Putative metal binding sites ..... 118
3.4.2.7 Highly coordinated water molecules ..... 120
3.4.2.8 The effects of crystal packing ..... 121
3.4.3 Comparison of the structures of Piwi and Ago Argonaute PAZ domain with RNA complexes ..... 123
3.4.3.1 Overall structure topology ..... 124
3.4.3.2 The binding pocket ..... 126
3.4.4 Implication for other Argonautes ..... 128
**3.5 Biochemical Analysis of the PAZ domain from the Piwi Subfamily protein Aubergine ..... 131**
3.5.1 Introduction ..... 131
3.5.2 Biophysical and biochemical studies of the Aubergine PAZ domain ..... 132
3.5.2.1 Analytical Gel Filtration ..... 132
3.5.2.2 Analytical Ultracentrifugation ..... 135
3.5.3 Interaction between Aubergine PAZ domain and methylated and unmethylated small RNA ..... 136
3.5.3.1 By Isothermal Titration Calorimetry (ITC) ..... 136
3.5.3.2 By Band Shift Assay ..... 137
3.5.3.3 By Flourescence Anisotropy ..... 138
3.5.4 Production of AubPAZ3$_{(279-398)}$ mutants ..... 139

3.5.4.1 Production of binding pocket mutant ..... 140
3.5.4.2 Production of SumoPAZ ..... 141
3.5.4.3 Production of SumoPAZ mutants ..... 142
3.5.5 Discussion ..... 146

**4 Discussion and Concluding remarks ..... 148**

**5 References ..... 158**

**Appendix ..... 173**

List of Figures


**Figure 1** The phylogenetic tree of Argonaute proteins ..... 4
**Figure 2** General Domain Structure of Argonaute proteins ..... 5
**Figure 3** Crystal structure of the tertiary complex ..... 6
**Figure 4** Summary of small RNA biogenesis pathways ..... 8
**Figure 5** The typical 3'- 2nt - overhang and 5' - phosphate group structure for miRNA duplex ..... 9
**Figure 6** Biogenesis pathway of C elegans or plant for siRNA ..... 12
**Figure 7** Schematic representations of piRNA biogenesis Pathways ..... 20
**Figure 8** Domain structure of subfamilies of Hen1 proteins ..... 27
**Figure 9** 3'piRNA end methylation by methyltransferase Hen1 in the presence of co - factor SAM ..... 29
**Figure 10** Interactions between 3'-end piRNA and the PAZ domain of Piwi subfamily Argonaute Proteins ..... 30
**Figure 11** Sequence alignment of full length animal Hen1 proteins ..... 51
**Figure 12** Small scale solubility testing for Hen1 proteins across human, *Drosophila* and zebrafish ..... 53
**Figure 13** Schematic prediction of the disordered regions in dHen1 ..... 55
**Figure 14** Schematic diagram of C-terminal truncated dHen1 constructs boundaries ..... 56
**Figure 15** C-terminal tail of dHen1 ..... 56
**Figure 16** SDS – PAGE gel of the soluble fractions for dHen1.4$_{(1-356)}$ expression test ..... 59
**Figure 17** SDS – PAGE gel result for step-wise purification scheme of dHen1.4$_{(1-356)}$ ..... 61
**Figure 18** dHen1.4$_{(1-356)}$ Gel filtration trace with SDS -PAGE gel of the eluted fractions ..... 62
**Figure 19** Result of a tryptic digest reaction of dHen1.4$_{(1-356)}$ ..... 63
**Figure 20** Schematic representation of the tryptic sensitive loop on dHen1.4$_{(1-356)}$ ..... 63
**Figure 21** The results of trypsin and subtilisin digest of dHen1.2$_{(1-391)}$ ..... 65
**Figure 22** Blot for N-terminal sequencing of trypsin and subtilisin digest product of dHen1.2$_{(1-391)}$ ..... 66
**Figure 23** The C-terminal sequence of dHen1 from residues 357 to 391 with secondary structure prediction by JPred ..... 67
**Figure 24** Final purification step - gel filtration traces and the SDS-PAGE results of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-164)}$ ..... 68
**Figure 25** Comparing the sizes of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ ..... 70
**Figure 26** Crystal structure of the Arabidopsis HEN1 ..... 73
**Figure 27** Schematic representations of the in-vitro methylation assay and detection by autoradiography ..... 76
**Figure 28** Domain structure of dHen1.4$_{(1-356)}$ as ΔC1 and dHen1.12$_{(1-368)}$ as ΔC5 ..... 77
**Figure 29** dHen1 methylates single-stranded RNAs at 26°C and 37°C ..... 78
**Figure 30** dHen1 FL and truncated constructs methylate single -stranded RNAs at 37°C ..... 80
**Figure 31** Schematic representations of the in *vitro* methylation assay and detection by liquid scintillation counting ..... 82
**Figure 32** Result of the liquid scintillation counting using two dHen1 constructs against 5 different lengths of small RNA substrates ..... 83
**Figure 33** Analytical Gel filtration ..... 84

**Figure 34** Trace of analytical gel filtration for dHen1.2$_{(1-391)}$, dHen1.4$_{(1-356)}$, dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ ..... 85
**Figure 35** Calculation of molecular weight from the elution volume based on the calibration ..... 86
**Figure 36** Schematic summary of the working hypothesis for the Hen1 protein mechanism ..... 89
**Figure 37** Schematic summary of currently available Argonaute protein structures ..... 91
**Figure 38** Sequence alignment of Argonaute protein PAZ domains across different subfamilies ..... 92
**Figure 39** Aubergine PAZ domain constructs summary ..... 93
**Figure 40** AubPAZ3$_{(279-398)}$ purification scheme ..... 95
**Figure 41** Mass spectrometry results confirms the molecular weight of the sample corresponds to that of AubPAZ3$_{(279-398)}$ ..... 96
**Figure 42** AubPAZ3$_{(279-398)}$/ RNA1 complex Crystals clusters in Natrix Screen ..... 98
**Figure 43** Crystals in 50mM MES pH6.1 ..... 99
**Figure 44** Schematic representations of the topology of the Aubergine PAZ domain ..... 106
**Figure 45** Structure of Aubergine PAZ domain in complex with the 7mer 2'-O-methyl RNA (5'-AACCUAmC-3') ..... 107
**Figure 46** Sequence alignment of PAZ domain for the Piwi subfamily of Argoanute proteins ..... 108
**Figure 47** The overall structure of binding pocket of Aubergine PAZ domain ..... 109
**Figure 48** Top down view of the Aubergine PAZ domain RNA binding pocket ..... 110
**Figure 49** Two molecules in the asymmetric unit of the crystal structure ..... 111
**Figure 50** Electron density at the 3'- end of the RNA (contoured at 1.5 σ) ..... 112
**Figure 51** Detailed recognition of 3'-end of piRNA by Aubergine PAZ domain RNA binding pocket ..... 113
**Figure 52** Charged surface representation of the Aubergine PAZ domain protein bound to the 7mer RNA with 2'-O-methyl modification at the 3'-end ..... 115
**Figure 53** Electron density for the 'flipped' base nucleotide ..... 116
**Figure 54** The 'flipped' nucleotide ..... 117
**Figure 55** Sequences conservation mapped onto the surface of Aubergine PAZ domain ..... 118
**Figure 56** Electron density representation of putative $Mg^{2+}$ ion stabilising the backbone of the RNA molecule ..... 119
**Figure 57** Highly coordinated water molecule in the RNA binding pocket ..... 120
**Figure 58** Molecule A and Molecule B with symmetry related partners ..... 121
**Figure 59** Symmetry related RNA inter-molecular interactions ..... 122
**Figure 60** Overall structural topology of Piwi subfamily (Aubergine) ..... 124
**Figure 61** Superposition the structures of hAgo1 PAZ domain in red bound to non-methylated small RNA in blue ..... 125
**Figure 62** Superimposed RNA binding pocket of Aubergine (shown in pink) and hAgo1 PAZ domain (shown in light blue) ..... 126
**Figure 63** Van der Waals radius representations of Aubergine and hAgo1 PAZ ..... 127
**Figure 64** Van Der Waals radius representations of the dAgo2 PAZ domain ..... 130
**Figure 65** Van Der Waals radius representations of the *Pyrococcus furiosus* Argonaute PAZ domain ..... 131

**Figure 66** Calibration run of the analytical gel filtration column superdex S75 ..... 133
**Figure 67** The calculations of the molecular weight from the elution volume based on the calibration ..... 134
**Figure 68** AUC raw data ..... 135
**Figure 69** EMSA result for AubPAZ3$_{(279-398)}$ with RNA2 (5'-AACCUAC-3') ..... 137
**Figure 70** EMSA result for AubPAZ3$_{(279-398)}$ at a higher concentration ..... 138
**Figure 71** Fluorescence anisotropy results for interactions between AubPAZ3$_{(279-398)}$ and RNA3 or RNA4 respectively ..... 139
**Figure 72** The three key residues in the binding pocket of Aubergine PAZ domain ..... 140
**Figure 73** Gel filtration trace and SDS-PAGE of SumoPAZ ..... 142
**Figure 74** Expression test for the SumoPAZ mutant ..... 146
**Figure 75** Summary of the structures for Piwi subfamily PAZ domains ..... 153
**Figure 76** Superimposed PAZ domains ..... 154
**Figure 77** RNA trajectories in the Piwi subfamily of PAZ domains ..... 155

List of Tables


**Table 1** List of primers and constructs for Hen1 protein expression .....32
**Table 2** List of primers and constructs for Aubergine PAZ domain proteins .....33
**Table 3** List of constructs and primers for the Aubergine PAZ domain protein mutants .....34
**Table 4** List of single stranded RNA sequences used with Hen1 proteins ...41
**Table 5** List of sequences of single stranded RNAs used with Aubergine PAZ domain proteins .....41
**Table 6** List of reagents used in Western blot .....45
**Table 7** Solubility testing results for dHen1 constructs .....57
**Table 8** List of expression conditions combinations for dHen1.4$_{(1\text{-}356)}$.....58
**Table 9** List of protein standards for gel filtration column calibration .....86
**Table 10** Analytical gel filtration results .....87
**Table 11** Aubergine PAZ domain constructs expression and solubility testing .....94
**Table 12** Crystallographic data and refinement statistics for AubPAZ3$_{(279\text{-}398)}$/RNA1 complex ..... 103
**Table 13** List of protein standards for gel filtration column calibration..... 133
**Table 14** Summary of the expression test for the SumoPAZ mutants ..... 145

# 1. Introduction

## 1.1. Overview of RNA silencing

RNA silencing is a form of post- transcriptional gene regulation. It was first discovered in the late 1990s by Andrew Fire and Craig Mello that, in *C. elegans,* double-stranded RNA (dsRNA) but not single-stranded RNA (ssRNA) can trigger gene silencing (Fire et al, 1998), a discovery for which they were awarded the Nobel Prize in 2006. One year later this phenomenon was also observed in plant and small interfering RNA (siRNA) was identified to be responsible for this post-transcriptional gene silencing activity (Hamilton & Baulcombe, 1999). Ever since then the field of RNA silencing has undergone astonishing development: small RNAs have been discovered across different species: animal, plants, fungi; many components of the various silencing effector complexes have also been identified (reviewed in (Ghildiyal & Zamore, 2009)). As we gain more understanding in this field, different sub-classes of small RNAs and their functioning pathways have emerged. Currently, according to the biogenesis and functions of the small RNAs, they can be broadly divided into three sub-classes: siRNAs, micro RNAs (miRNAs) and Piwi interacting RNAs (piRNAs).

The overall RNA silencing event can be summarised as sequential steps of: small RNA production and biogenesis, sorting and incorporation of small RNAs into Argonaute proteins, target recognition and biological effector function.

### 1.1.1. Small non-coding RNAs as regulator molecules

The first small RNA, *lin-4*, was discovered in 1993. It contains sequence complementarities to the mRNA of *lin-14* and represses the expression of *lin-14*

protein in *C. elegans* (Lee et al, 1993) (Wightman et al, 1993). Owing to the development in cloning techniques and deep sequencing technologies (Lu et al, 2005) (Lai et al, 2003), the number of known small non-coding RNAs now has expanded dramatically. Found in eukaryotes, these small RNAs regulate gene expression through a variety of methods including translational repression, mRNA degradation and heterochromatin formation (Lagos-Quintana et al, 2001) (Lee & Ambros, 2001) (Lau et al, 2001) (Elbashir et al, 2001) (Reinhart & Bartel, 2002) (Hannon, 2002) (Kim et al, 2008). They function in a range of different biological processes including genome surveillance, cell development and differentiation, transposon silencing and antiviral defences (Filipowicz et al, 2008) (Vagin et al, 2006).

The defining features of these non- coding small RNAs are that they are only between 20 – 30 nucleotides in length and they associate with a member of the Argonaute protein family to drive the regulation of gene expression (Zamore et al, 2000) (Hammond et al, 2000). The small RNAs provide the specificity via base pairing and act as a guide to the Argonaute proteins and direct them to the target molecules. This RNA-Argonaute complex forms the core of the RNA silencing machinery RNA Induced Silencing Complex (RISC), which recognises the target mRNA and represses gene expression (Miyoshi et al, 2005) (Rand et al, 2005) (Matranga et al, 2005).

### 1.1.2. Argonaute proteins – the common theme

Argonaute proteins were originally discovered from developmental studies in plant and were named after the knockout phenotype in *Arabidopsis thaliana* for its resemblance of the tentacles of the octopus *Argonauta argo* (Bohmert et al,

1998). Later on, they were also found to be important in germ cell division in *Drosophila* (Cox et al, 1998). Since the discovery of RNA silencing, Argonaute proteins have been found to incorporate small RNAs to form the effector complexes in RNA silencing and have been identified as the principle protein component of the silencing machinery through genetic and biochemical studies (Meister et al, 2004) (Liu et al, 2004) (Song et al, 2004) (Carmell et al, 2002).

As shown in a reconstitution study, the minimal RISC (RNA Induced Silencing Complex, the effector complex of RNA interference) contains an Argonaute protein bound to a small RNA. Thus the Argonaute-RNA complex is sufficient to direct silencing. However, in *vivo* it is likely that Argonaute proteins function in larger complexes and interact with multiple proteins at different stages of RNA silencing pathways (Rivas et al, 2005).

Argonaute proteins are found across a very board range of species including mammals, *Drosophila*, *S. pombe*, *C. elegans*, plants and some prokaryotes. Though the substrates and the biological functions of the prokaryotic Argonaute proteins are unknown. Argonautes can be subdivided into several major subfamilies. **Figure 1** shows the divisions of different Argonaute protein subfamilies (Parker & Barford, 2006).

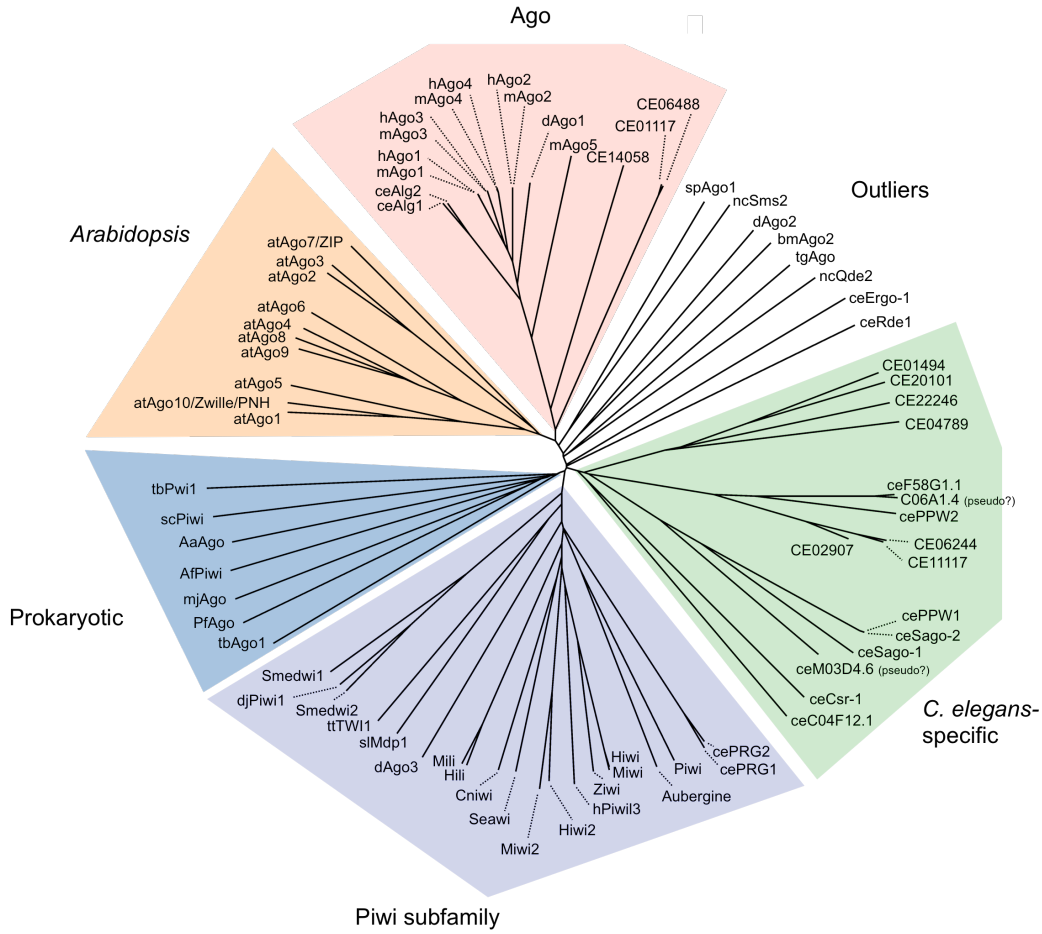



**Figure 1: The phylogenetic tree of Argonaute proteins shows that the Argonaute proteins are broadly divided into several subfamilies of Ago, Piwi, *Arabidopsis* (Plant), prokaryotic (Bacteria), C. *elegans* specific (Worm) and outliers (Parker & Barford, 2006).**

In most animals including humans, the Argonaute proteins fall broadly into two sub– classes: the **Ago subfamily** with proteins similar to *Arabidopsis* AGO1 and the **Piwi subfamily** with proteins similar to the *Drosophila* Piwi protein shown in **Figure 1**. Importantly, different subclasses associate with different forms of small RNA. Ago Argonautes associate with siRNAs and miRNAs and act on downstream target molecules in the siRNA pathway and the miRNA pathway respectively; Piwi Argonaute proteins bind to piRNAs and are involved in the piRNA pathway.

The number of Argonaute proteins present in each species varies significantly. For example, in human there are eight Argonaute proteins, in *C. elegans* there are at least twenty six Argonaute proteins and in *Arabidopsis* there are ten (Tolia & Joshua-Tor, 2007). Argonaute proteins have also been found in prokaryotes, eubacteria and achaebacteria, implying that they are evolutionary ancient, though the functions of prokaryotic Argonautes are still unclear.

### 1.1.3. Overview of Argonaute structure and mechanism

The Argonaute proteins share a similar overall domain architecture as shown in **Figure 2A**.

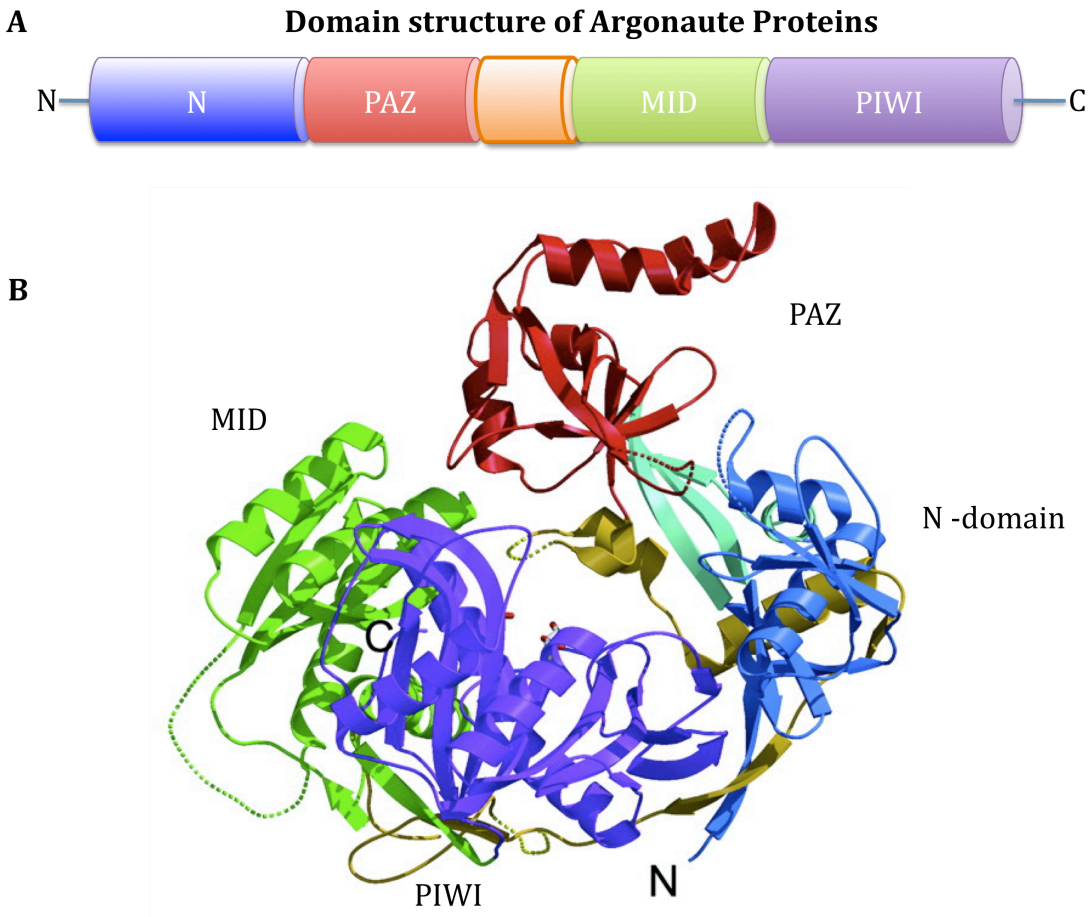



**Figure 2: A) General domain structure of Argonaute proteins. B) Crystal structure of the full length prokaryotic (*Pyrococcus furiosus*) Argonaute protein. Each domain is coloured corresponding to the domain structure in Figure 2A. Figure from (Song et al, 2004).**

The majority of the structural studies so far on Argonaute proteins have been conducted on prokaryotic Argonaute proteins (Wang et al, 2009b) (Yuan et al, 2006) (Parker et al, 2005) (Song et al, 2004). The first crystal structure of a full length Argonaute protein solved in 2004 was the Argonaute protein from the bacterium *Pyrococcus furiosus,* the overall structure as shown in **Figure 2B** (Song et al, 2004). In 2009, the crystal structure of full length *Thermus thermophilus* Argonaute protein in complex with nucleic acid guide and target strands was solved (Wang et al, 2008) (**Figure 3**).

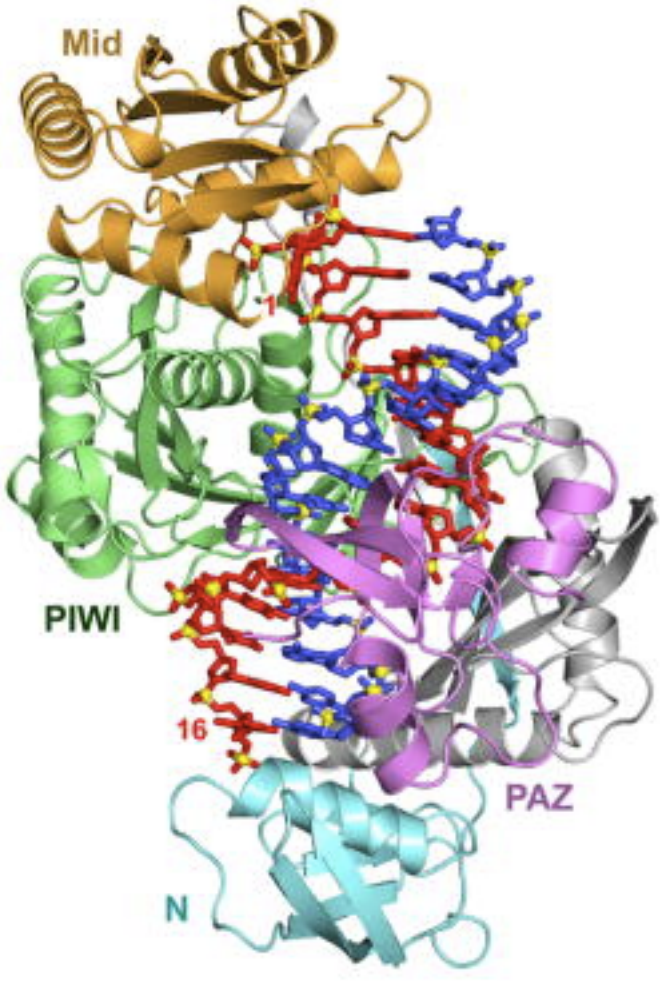



**Figure 3: Crystal structure of the tertiary complex comprising of *Thermus thermophilus* Argonaute protein in complex with a DNA guide strand (red) and RNA target strand (blue) (Wang et al, 2009b). Prokaryotic Argonaute, for reasons unclear, seem to prefer short DNA guide.**

Together the various structural studies have shown that the Argonaute protein is a bi-lobal protein. The N-terminal lobe consists of the N-domain and PAZ domain linked by a flexible linker region. The C-terminal lobe consists of MID and PIWI domains. The PAZ domain contains a binding pocket for the 3'-end of the guide strand RNA (DNA in the case of prokaryotes) (Lingel et al, 2003) (Song et al, 2003). MID and PIWI domains form the binding site for anchoring the 5'-phosphate of the guide strand RNA (Parker et al, 2004) (Ma et al, 2005). The PIWI domain has an RNase-H-like fold and some Argonaute PIWI domains employ similar activity to RNase-H-like processing, termed "slicing" (Song et al, 2004). Other Argonaute proteins that lack the slicing activity usually involve other protein factors to repress the target molecules. The "slicing" catalytic site of the Argonaute protein locates in the PIWI domain with a conserved DDH (Asp-Asp-His) motif for divalent metal binding. Slicing activity generates products containing 3'-OH group and 5'-phosphate group.

### 1.1.4. Three major classes of small RNA silencing pathway

The small RNAs are broadly divided into three major classes by their biogenesis, mode of action and their association with a member of the Argonaute protein family. The individual combination of a small RNA and an Argonaute protein determines its biological function pathway. The schematic diagram of each pathway is summarised in **Figure 4** and described briefly in the next three sections:

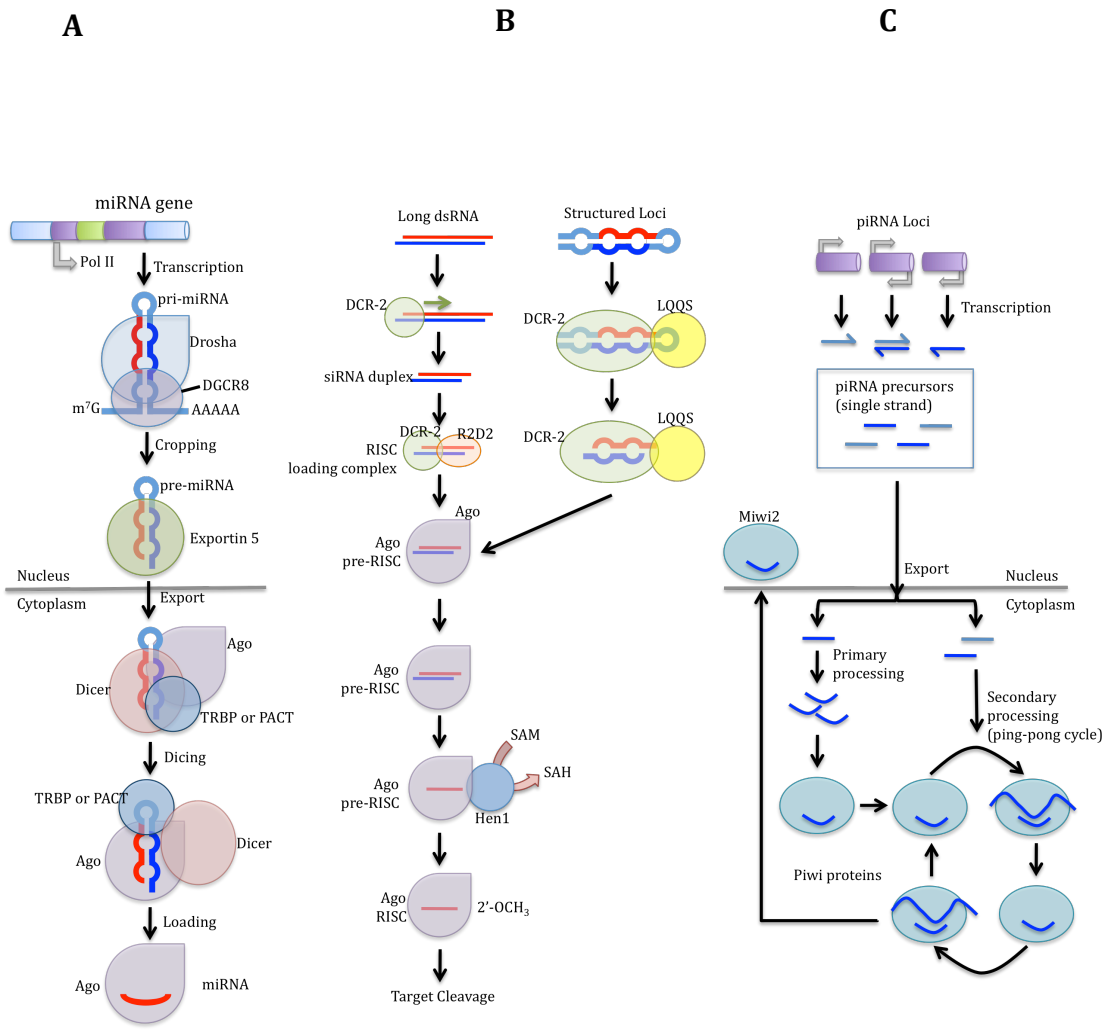



**Figure 4: Summary of small RNA biogenesis pathways. A) miRNA biogenesis pathway B) siRNA biogenesis pathway in *Drosophila* C) piRNA biogenesis pathway**

#### 1.1.4.1. miRNAs

The miRNA pathway is probably the best understood pathway compared to the siRNA and piRNA pathways. Mature miRNAs are approximately 22 nucleotides in length and are bound to the Ago subfamily Argonaute proteins (Kim, 2005). miRNA target mRNAs through base pairing in the 3'-untranslated region (3'-UTR) and result in mRNA destabilisation and

translational repression (Eulalio et al, 2008). The canonical miRNA biogenesis pathway is shown in schematic diagram **Figure 4A**.

Most miRNAs are first transcribed by DNA-dependant-RNA-Polymerase II from the genome and the transcripts form the Pri-miRNAs containing 5'-m7G Cap and 3'-polyA tail features (Lee et al, 2004a) (Cai et al, 2004). Pri-miRNA then associates with the double-stranded RNA binding protein DiGeorge Syndrome critical region gene 8 (DGCR8) and RNase III enzyme Drosha to form the Microprocessor complex. The cleavage product of Drosha is a 60 – 70 nt stem-loop precursor of miRNA termed Pre-miRNA with a 2 nt-overhang at the 3'-end (Lee et al, 2003) (Denli et al, 2004) (Han et al, 2004) . The Pre-miRNA is then recognised by the nuclear transporter protein Exportin-5 and transported into cytoplasm (Yi et al, 2003) (Lund et al, 2004) (Okada et al, 2009).

Once in the cytoplasm the Pre-miRNA is further processed by a protein complex composed of another RNase III enzyme Dicer, double-stranded RNA (dsRNA) binding protein TAR RNA-binding Protein (TRBP) and PACT into a 22 nt miRNA duplex containing a guide and a passenger strand with 5'-phosphate group and 3'-dinucleotide overhang shown in **Figure 5** (Bernstein et al, 2001) (Hutvagner et al, 2001) (Macrae et al, 2006).

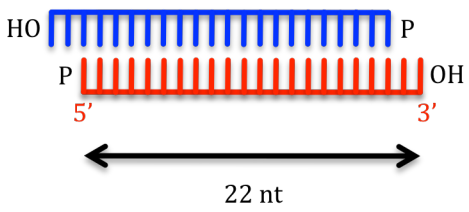



**Figure 5: The typical 3'- 2 nt - overhang and 5' - phosphate group structure for miRNA and siRNA duplex.**

Upon cleavage, the 22 nt miRNA duplex is loaded into an Argonaute protein by unknown mechanisms. Together Dicer, Argonaute and TRBP/PACT protein complex form the RISC Loading complex (RLC) (Chendrimada et al, 2005). The passenger strand in the miRNA duplex is discarded and degraded. The guide strand miRNA then guides the matured RISC complex to its molecular target (MacRae et al, 2008) (Gregory et al, 2005).

In plant, after the miRNA duplex is generated by Dicer, they are methylated on the 2'-OH at the 3'-end to mark the maturation of miRNA and prevent it from degradation (Li et al, 2005). This additional modification step is carried out by the methyltransferase HUA ENHANCER1 (HEN1) (Yu et al, 2005). Similar modifications are found on piRNAs and *Drosophila* siRNAs and will be discussed in more detail in Section 1.2, Section3.3 and Section3.4.

#### 1.1.4.2. siRNAs

siRNAs are approximately 21 nt long small RNAs that were first discovered in plants (Hamilton & Baulcombe, 1999). The siRNAs are derived from dsRNA precursors and they associate and utilise the slicing activity of the Ago subfamily of Argonaute proteins for target cleavage through complementary base pairing (Kawamura et al, 2008) (Czech et al, 2008). The dsRNA precursor of siRNAs can be endogenously produced (endo-siRNAs) or exogenously introduced (exo-siRNAs). The endogenous siRNAs precursors are predominantly derived from transposon transcripts, sense and antisense transposon transcript pairs with hairpin structures (Babiarz et al, 2008) (Okamura et al, 2008) (Ghildiyal et al, 2008). The exogenous siRNAs are mostly derived from viral infection or experimentally introduced long dsRNAs or siRNA duplexes.

As a model system, the siRNA pathway is best understood in *Drosophila*. The Schematic diagram of the siRNA biogenesis is summarised in **Figure 4B**.

In *Drosophila*, long dsRNA precursors, either transcribed from the genome or as a product of viral infection, are subjected to sequential cleavage by an RNase III enzyme Dicer 2 (Lee et al, 2004b) (Liu et al, 2003) in complex with the double-stranded RNA binding Domain (dsRBD) proteins LOQS (Zhou et al, 2009) (Hartig et al, 2009) (Forstemann et al, 2005) and R2D2 (Liu et al, 2003) (Marques et al) (Czech et al, 2008) (Saito et al, 2005) (Brennecke et al, 2007). The cleavage products of Dicer form the siRNA duplex that is similar to the miRNA duplex (**Figure 5**), is then loaded onto Ago2 (an Ago subfamily Argonaute protein) in the RISC complex (Kawamura et al, 2008). The passenger strand in the siRNA duplex is cleaved and discarded in the matured RISC. A further step in *Drosophila* but not mammal is required for full RISC maturation: the 2'-O-methylation at the 3'-end of the single stranded guide RNA by methyltransferase Hen1, the analogue of plant HEN1 described in the previous section (Horwich et al, 2007).

In *C. elegans* and plant, the production of siRNAs can also involve the activity of RNA-dependant RNA Polymerases (RdRPs) but in different stages as summarised in **Figure 6.**

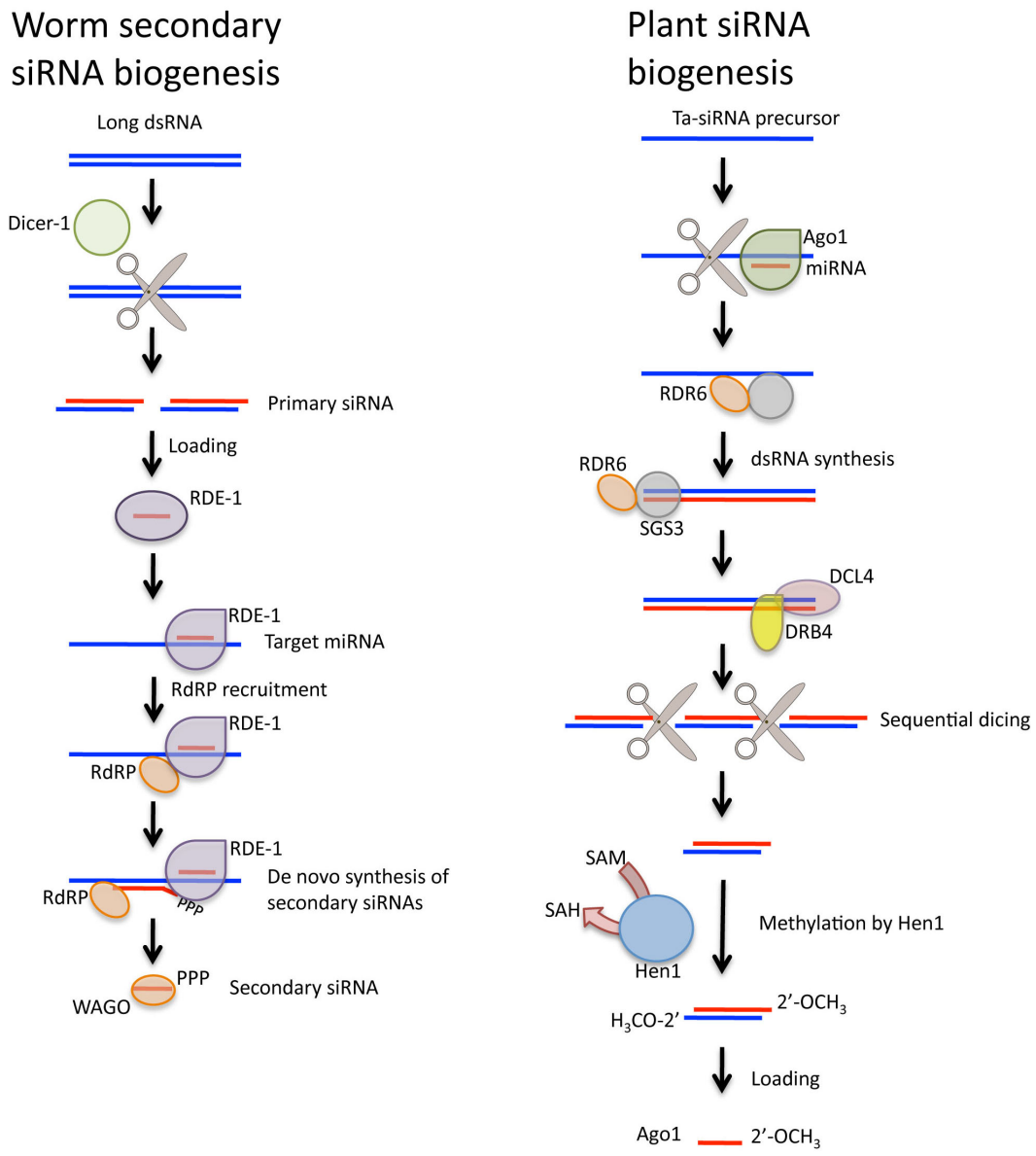



**Figure 6: Biogenesis pathways of *C. elegans* and plant siRNAs**

In *C. elegans* the primary siRNAs are produced from long dsRNA precursors in the same manner as in *Drosophila* involving Dicer but not RdRPs. The primary siRNAs then guide the Argonaute protein (RDE-1) to the mRNA targets and recruit RdRPs to synthesise directly the secondary siRNAs using the mRNA targets as the template without the involvement of Dicer (Steiner et al, 2009). The secondary siRNA have a 5' – triphosphate group, which is consistent with

the synthesis being carried out by a polymerase. The secondary siRNAs can then guide the *C. elegans*-specific subfamily Argonaute protein (WAGO) (**Figure 1**) for further amplification of the silencing process (Pak & Fire, 2007) (Sijen et al, 2007).

Plant siRNAs are further divided into three subclasses: the ta-siRNAs (tans-acting siRNAs), nat-siRNAs (natural antisense transcript-derived siRNAs) and heterochromatic siRNAs (hcsiRNAs). The overall biogenesis of plant siRNAs first involves converting single stranded precursors to double stranded RNA by RdRPs. The products of the RdRPs are then processed by Dicer and loaded onto Argonaute proteins as effector complex in a similar manner to that in *Drosophila* (Herr & Baulcombe, 2004) (Chapman & Carrington, 2007). The difference is that, during the siRNA biogenesis, the siRNA duplex is methylated by HEN1 on both strands *before* loading onto the Argonaute proteins in the final effector complex; whereas in *Drosophila* only the single stranded siRNAs are methylated *after* the passenger strands are discarded in RISC. Each subclass of the siRNAs are produced involving a distinct member of Dicer family protein and are loaded to a specific Argonaute protein in the effector complex. The ta-siRNAs are derived from the TAS transcripts; nat-siRNAs are produced often when the plant is under biotic and abiotic stress and the hcsiRNAs are derived from repeats and transposable elements (Borsani et al, 2005) (Katiyar-Agarwal et al, 2006) (Henz et al, 2007).

#### 1.1.4.3. piRNAs

piRNAs are the most recently discovered subset of small RNAs (Aravin et al, 2001) (Aravin et al, 2003). They are germ cell specific and they associate with

the Piwi subfamily of Argonaute proteins (Gunawardane et al, 2007) (Nishida et al, 2007) (Aravin et al, 2006) (Girard et al, 2006) (Grivna et al, 2006a). They are longer than siRNAs or miRNAs at approximately 28 nt in length. They are found in *Drosophila* (where most of the studies on piRNAs have been done), mouse, zebrafish, worm and human but not in plant or bacteria. The one key known role of piRNAs is to control against mobile genetic elements (Aravin et al, 2003) (Houwing et al, 2007) (Aravin et al, 2006) (Girard et al, 2006) . The piRNAs are believed to evolve to protect the genome from the activity of parasitic nucleic acids, and preserve germline genome integrity and stability by repressing transposons (Aravin et al, 2007a) (Saito et al, 2006) (Sarot et al, 2004) (Vagin et al, 2006). They are also important in germ cell development (Carmell et al, 2007) (Houwing et al, 2007).

The defining structure feature for piRNAs is that they are ubiquitously methylated at the 3'-end on the 2'-OH group by Hen1 protein in a similar manner to that in *Drosophila* siRNAs (Ohara et al, 2007) (Kirino & Mourelatos, 2007a) (Vagin et al, 2006). In addition piRNAs differ from siRNAs and miRNAs in several aspects including their biogenesis. The main focus of this study is on piRNAs, which are described in more detail in the next section.

## 1.2.piRNAs

### 1.2.1. The discovery of piRNAs

piRNAs were discovered as a result of small RNA profiling studies in *Drosophila*. They were originally termed repeat-associated-small-interfering RNA (rasiRNA) as they corresponded extensively with repeat elements and are involved in the silencing of transposons in the germline (Aravin et al, 2001) (Aravin et al, 2003).

Further studies confirmed that they are distinct from the siRNAs and miRNAs as a separate subclass of small RNAs: they are longer in length (22-29 nt long) (Aravin et al, 2003); they associate with the Piwi subfamily of Argonaute proteins as opposed to the Ago subfamily of Argonaute proteins which is the case for siRNAs and miRNAs; they also do not require the involvement of RNase III enzyme Dicer for their biogenesis. In addition, they are ubiquitously methylated on the 2'-OH group at the 3'-end by Hen1 protein, similar to plant siRNAs, miRNAs and siRNAs in *Drosophila* but not any other small RNAs species in animal (Vagin et al, 2006) (Saito et al, 2006) (Girard et al, 2006) (Grivna et al, 2006a) (Houwing et al, 2007). They therefore were renamed piRNAs (Piwi interacting RNAs).

piRNAs have extremely diverse sequences. So far more than 1.5 million piRNA sequences have been identified in *Drosophila* compared to in the order of hundreds for miRNAs. They do not possess any sequence motif except for a strong bias for Uridine at position 1 of the 5'-end. Despite the large number of piRNAs and their sequence diversity, they can be mapped on to only a few hundred genome regions, termed piRNA clusters (Czech et al, 2008) (Kawamura et al, 2008) (Chung et al, 2008) (Saito et al, 2006) (Ruby et al, 2007) (Brennecke et al, 2007) (Nishida et al, 2007) (Malone et al, 2009). The best characterised piRNA cluster is the *flamenco* locus which encodes piRNAs that are anti-sense to the *gypsy*, *ZAM* and *Idefix* retroviral elements in *Drosophila* (Pelisson et al, 2007) (Prud'homme et al, 1995) (Pelisson et al, 1994) (Desset et al, 2008) (Mevel-Ninio et al, 2007).

In addition to the piRNAs derived from repeat sequences in the genome, there is a subclass of piRNAs that are not derived from the repetitive sequences (Watanabe et al, 2008) (Houwing et al, 2007) (Aravin et al, 2006) (Lau et al, 2006) (Aravin et al, 2007b) (Girard et al, 2006) (Grivna et al, 2006a) (Grivna et al, 2006b). In mammals, piRNAs can be further divided into pre-pachytene piRNAs that are produced before the meiotic pachytene, a stage during Prophase I of meiosis where chromosomal crossover occurs through homologous recombination between the non-sister chromatids; and pachytene piRNAs that are produced during the pachytene stage of meiosis in germ cell development. Unlike pre-pachytene piRNAs, most of the pachytene piRNAs are derived neither from the repeat elements nor transposons and their functions are still unclear (Aravin et al, 2007b) (Girard et al, 2006) (Kuramochi-Miyagawa et al, 2008) (Aravin et al, 2008). In *C. elegans*, the piRNAs are termed 21U RNAs for they are exactly 21 nucleotides in length with a 5'-Uridine on position one. Although they associate with the C. *elegans* Piwi proteins PRG1 and PRG2 and are required for germline maintenance, their targets also remain elusive (Ruby et al, 2006) (Batista et al, 2008) (Das et al, 2008) (Wang & Reinke, 2008).

### 1.2.2. Functions of piRNAs

The main known function of piRNAs is to repress transposable elements in the germ line, thus maintaining genome integrity. Germ cell genomic DNA forms the basis for inheritance, so it must be subject to greater protection than DNA from the uni-generational somatic cells. Transposable elements are found across many different species and they make up a significant proportion of the genome. In human, around 45% of the genome is comprised of transposable elements

(Lander et al, 2001) and 15% – 22% of the *Drosophila* genome is made up of transposable elements too (Kapitonov & Jurka, 2003) (Biemont & Vieira, 2006). Although transposable elements can have occasional beneficial biological implications such as protection of the telomeric ends of the linear chromosomes in *Drosophila* (Pardue et al, 2005), the majority of the transposable elements are considered to be parasitic and their mobile nature can lead to genome instability through structural re-arrangement such as duplication and deletions (reviewed in (Goodier & Kazazian, 2008) (Gogvadze & Buzdin, 2009). Subsequently many organisms have evolved endogenous pathways such as the piRNA silencing pathway against the transposable elements activities (Kazazian, 2004).

piRNAs are predominately found in the germ cells. In animal, transposable elements are also silenced by endo-siRNAs, which are generated endogenously from transposable elements, RNA hairpin structure loci, but more predominantly in *soma* as opposed to in the germ line. Somatic piRNAs are believed to be produced when mutations disrupt the endo-siRNAs pathway and to rescue transposon suppression (Ghildiyal et al, 2008) (Kawamura et al, 2008).

In the piRNA silencing pathway, piRNAs guide a member of the Piwi Argonaute protein subfamily to silence the complementary target RNAs such as transposon RNAs. In *Drosophila*, the three Piwi Argonaute proteins (Aubergine, Piwi and Ago3) act non-redundantly in suppressing transposons in the germ line. Mutation of these Piwi subfamily of Argonaute proteins result in developmental defects and infertility (Li et al, 2009) (Cox et al, 1998) (Lin & Spradling, 1997) (Schmidt et al, 1999) (Cox et al, 2000). In mouse, knock out of Mili and Miwi2 (but not Miwi) Piwi subfamily of Argonaute proteins result in gametogenesis

arrest and sterility in males. The defects are thought to be the result of retrotransposon gene expression activation through impaired DNA methylation in the regulatory region of the retrotransposons with the loss of Mili and Miwi2. Double-stranded DNA breaks are generated during transposable elements transposition, which activates a DNA damage checkpoint. (Carmell et al, 2007) (Aravin et al, 2007b) (Kuramochi-Miyagawa et al, 2008) (Klattenhoff et al, 2007).

In addition, genetic studies suggest that, in mouse, Mili and Miwi2 in complex with piRNAs are not only involved in the post-transcriptional regulation of genes but also in guiding CpG methylation at the transposable elements loci, which leads to the transcriptional silencing of LINE (Long Interspersed Nuclear Elements) and LTR (Long Terminal Repeat) retrotransposons. The existence of transcriptional silencing involving chromatin by the piRNA pathway is also suggested in *Drosophila*, although the exact mechanisms remain unresolved (Aravin et al, 2007b) (Kuramochi-Miyagawa et al, 2008).

piRNAs are also suggested to target protein coding genes as a small proportion of piRNA sequences do not map onto any transposable elements sequence in the genome (Aravin et al, 2001). In *Drosophila*, piRNAs associated with Piwi proteins were shown to be involved in silencing *Stellate* and *Fasciclin 3* genes. *Stellate* gene located on the X chromosome is regulated by the piRNAs (*Suppressor of Stellate* tandem repeats) located on the Y chromosome. Deletion of suppressor of stellate tandem repeats results in overexpression of *Stellate* in *Drosohpila* testis that leads to the formation of crystalline aggregates containing *Stellate*-coding proteins, and this causes male sterility. *Fasciclin 3* gene encodes a cell adhesion molecule. Loss of Piwi protein in *Drosophila* results in over expression of

*Fasciclin 3*; the germ cells and somatic cells failing to intermingle and thus causing oogenesis defects (Saito et al, 2009) (Li et al, 2003) (Aravin et al, 2001) (Vagin et al, 2006) (Aravin et al, 2004). piRNAs associated with Aubergine were suggested to target the *vasa* gene. Overexpression of the *vasa* gene results in over sized testes when *vasa* transcripts are not silenced by the AT-chX piRNAs associated with Aubergine (Nishida et al, 2007) (Nagao et al, 2010). In mouse, the first piRNAs discovered were the pachytene piRNAs that did not target transposons (Aravin et al, 2007a). The lack of sequence complementarities between the piRNAs and the targets, the shortage of *in vivo* evidence of protein coding gene silencing by piRNAs marks the need for further investigation into the functions of piRNAs on targets other than transposable elements.

### 1.2.3. piRNA biogenesis

Mature piRNAs are derived from long single stranded RNA precursors transcribed from the piRNA clusters. This is in major contrast to the double-stranded precursors of siRNAs and miRNAs. Subsequently the long precursors are processed into mature piRNAs by two pathways, namely the primary processing pathway and the ping-pong amplification cycle, which are conserved across many animal species (Lau et al, 2006) (Houwing et al, 2008) (Kawaoka et al, 2009) (Robine et al, 2009). Both pathways generate effective mature piRNAs for transposon silencing and they also work synergistically in defending transposable element activity.

The primary processing pathway first generates a pool of piRNAs from genome to silence the target transposons; then it initiates the secondary pathway also known as the Ping-Pong cycle, to further amplify the population of piRNAs that

target the active transposons. The two pathways are summarised in the schematic diagram in **Figure 7**.

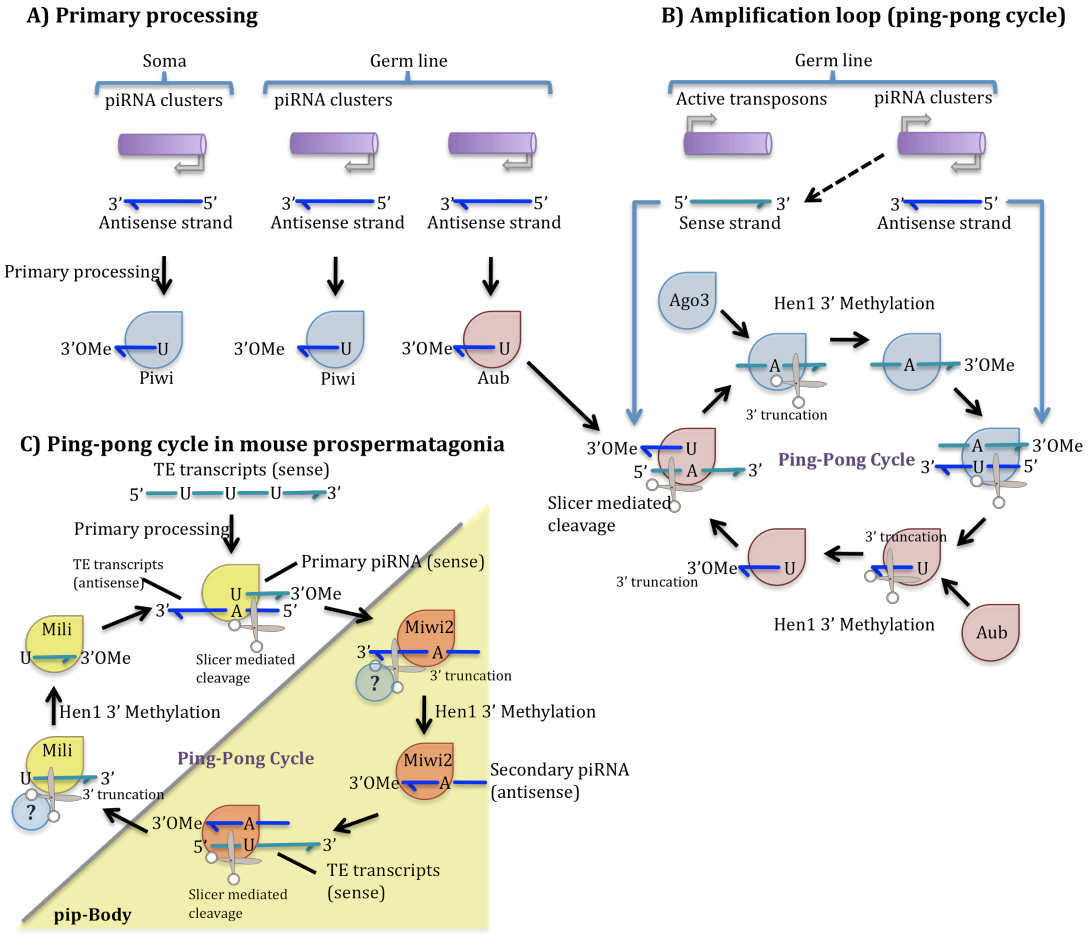



**Figure 7: Schematic representations of piRNA biogenesis Pathways. A) *Drosophila* Primary processing pathway B) *Drosophila* Ping-Pong amplification loop C) Mouse piRNA biogenesis Pathway**

### 1.2.3.1. Primary processing pathway

The genomic origins of the majority of the piRNAs are the piRNA clusters and each cluster encodes multiple precursor sequences of piRNAs. These clusters are rich in sequences of transposable elements, some clusters contain sequences encode for piRNAs on both strands (Aravin et al, 2008) (Brennecke et al, 2007); some clusters are transcribed only in one direction and the piRNAs sequences derived form these uni-directional clusters are often antisense to transposon transcripts (Girard et al, 2006) (Brennecke et al, 2007). It is shown that piRNAs

precursors are single stranded long RNAs and do not require the activity of RNase III enzyme Dicer for processing, but it is still unknown how primary piRNAs are generated from the piRNA clusters, possibly involving an unknown nuclease to generate primary piRNAs with 5' Uracil (Aravin et al, 2006) (Girard et al, 2006) (Aravin et al, 2007a) (Brennecke et al, 2007) (Malone et al, 2009). Several other proteins such as putative RNA helicase Armitage (ARMI), Female Sterile 1 YB (FS(1)YB) and putative nuclease Zucchini (ZUC) have recently been implicated in the biogenesis of the primary piRNAs in *Drosophila* (Saito et al, 2010) (Robine et al, 2009) (Haase et al, 2010) (Olivieri et al, 2010) (Qi et al, 2010).

**Figure 7A** shows the Primary piRNA processing pathway in *Drosophila*. Piwi is the only Piwi subfamily Argonaute protein that expresses in the ovarian somatic cells; therefore, the somatic piRNAs such as those derived from *flamenco* loci are bound exclusively with Piwi proteins (Desset et al, 2008) (Cox et al, 1998) (Cox et al, 2000) (Malone et al, 2009). In germ cells, after primary processing, the piRNAs that are antisense to the transposons are bound to the Aubergine and Piwi proteins. The mature primary piRNAs in both the germ cells and *soma* bear the 2'-O-methyl modification at 3'-termini and they are antisense to the transposon transcripts with a strong preference for U at the 5' termini.

The primary piRNA processing biogenesis pathway is the only means of generating piRNAs in certain cells such as *Drosophila* somatic follicle cells where only Piwi protein is expressed (Malone et al, 2009) (Saito et al, 2009) (Lau et al, 2006). In other cell types, such as *Drosophila* germ cells, the piRNA population is further amplified by the Ping-Pong cycle.

In mouse, the primary piRNAs are produced through unknown mechanisms from the long single stranded transposon mRNAs. Unlike the primary piRNAs in *Drosophila*, the mouse primary piRNAs are mostly sense strand of transposon transcripts and loaded onto the Mili Argonaute protein. They also bear the 2'-O-methyl modification at 3'-termini. The pre-pachytene piRNAs are then also further amplified through the Ping-Pong cycle. The pachytene piRNAs, on the other hand, mainly associate with Mili and Miwi and are thought be produced only through the primary processing pathway (Aravin et al, 2007b) (Aravin et al, 2008) (Girard et al, 2006).

#### 1.2.3.2. Ping-Pong amplification cycle

In *Drosophila*, it was found that Ago3 associates with piRNAs derived from the sense strand of transposon transcript, whereas the piRNAs that are typically bound to Piwi and Aubergine are antisense strands to the transposon mRNA (Saito et al, 2006) (Brennecke et al, 2007) (Nishida et al, 2007) (Yin & Lin, 2007). In addition, there is a striking sequence complementarity between the first 10nt of the sense piRNAs and antisense piRNAs (Gunawardane et al, 2007) (Brennecke et al, 2007). Because the slicer activity of Argonaute cuts the target across from nucleotide 10 of the guide strand, the complementarity between the sense and the antisense piRNAs over the first 10 nucleotides implicated Argonaute as the enzyme that generates the 5'-ends of the piRNAs. Together a "Ping-Pong" amplification model for piRNA biogenesis was proposed and is summarised in **Figure 7B**. In this model, the 5'-end of the piRNAs formation is the result of the Piwi subfamily of Argonaute (Aubergine and Ago3) proteins'

slicer activity; however the mechanism for generating the 3'-end of the piRNAs and the function of the 3'-end methylation, remained obscure.

In mouse similar biogenesis pathways for pre-pachytene piRNAs involving the Ping-Pong amplification loop are also present. The schematic of the mouse biogenesis pathway is summarised in **Figure 7C**.

As seen in **Figure 7C,** the mouse Ping-Pong amplification loop is very similar to that in *Drosophila*. The difference is that the mouse primary piRNAs tend to be the sense strand of transposon transcripts bound to Mili protein and the secondary piRNAs (piRNAs produced from the amplification loop) bound to the Miwi2 proteins tend to be anti-sense to the transposon transcripts. This preference of the strand orientation is the opposite to that of *Drosophila* (Aravin et al, 2008).

The biological implication of the primary and Ping-Pong amplification pathways was proposed to be an analogy of the adaptive immune system in that the primary processing pathways generates multiple piRNAs against the transposons and the Ping-Pong amplification loop amplifies the populations of piRNAs to target *active* transposons and mount a larger scale defence against the transposon activities (Aravin et al, 2008). In other words, production of RNAs such as the transcription of a transposon, which corresponds to one half of the Ping-Pong amplification cycle, initiates the Ping-Pong cycle and results in amplification of both sense and antisense strands of piRNAs.

### 1.2.4. piRNA regulation by Tudor-containing proteins

It has been recently discovered that Piwi Argonaute proteins in various species including *Drosophila*, mouse and *Xenopus laevis* contain symmetrical dimethylarginines (sDMA) in the arginine-glycine rich region at the N-terminus by immunoprecipitation (Vagin et al, 2009a) (Kirino et al, 2009a) (Nishida et al, 2009). This modification has provided a binding platform for the Tudor (Tud) domain containing proteins (the Tudor proteins), which have been implicated in gametogenesis and associate with the Piwi proteins via the sDMA (Vagin et al, 2009b) (Chen et al, 2009) (Reuter et al, 2009) (Shoji et al, 2009) (Vasileva et al, 2009) (Wang et al, 2009a) (Kirino et al, 2010). These arginine residues are methylated by Protein Arginine N-methyltransferase 5 (PRMT5) and are conserved in the Piwi subfamily of Argonaute proteins but are not found in the Ago subfamily of Argonaute proteins, suggesting that the modification is piRNA pathway specific.

In *Drosophila*, Tud, a 280kDa Tudor protein containing 11 Tudor domains, was shown to interact with Aubergine and Ago3 (Piwi subfamily proteins) in an sDMA – dependant manner. Structural studies have revealed that sDMA on the Aubergine protein were recognised by an asparagine gated aromatic cage formed by four aromatic residues in the Tudor domain of Tud (Boswell & Mahowald, 1985) (Liu et al, 2010a) (Liu et al, 2010b).

Similar interactions between Tud family proteins TDRD (Tud- domain containing) and the Piwi subfamily of Argonaute proteins in mouse have also been observed. TDRD6, homologue of *Drosophila* Tud, contains 7 Tudor domains and interacts with Mili and Miwi in an sDMA –dependant manner (Hosokawa et

al, 2007). Other TDRD proteins including TDRD1, TDRD2, TDRD4, TDRD7, TDRD8 and TDRD9 also demonstrated interactions with Piwi subfamily of Argonaute proteins (Chen et al, 2009) (Kirino et al) (Kojima et al, 2009) (Shoji et al, 2009) (Vagin et al, 2009b) (Vasileva et al, 2009) (Wang et al, 2009a) (Chuma et al, 2006). However it is unclear that some of these interactions, such as TDRD1 with Miwi and Mili or TDRD9 with Miwi2, are sDMA-dependant.

Together, this complex interaction and recognition between Tudor family proteins and the Piwi subfamily of Argonaute proteins suggest that Tudor proteins with multiple Tud domains could potentially interact with multiple Piwi subfamily of Argonaute proteins via their sDMA modifications and provide a binding scaffold for a multi-protein complex formation. This could be particularly relevant to the Ping-Pong amplification cycle, as it requires the activity of multiple Piwi subfamily Argonaute proteins simultaneously.

Loss of Tudor family proteins results in impaired piRNA production. Although the piRNAs were not abolished in Tudor mutants, piRNAs populations associated with the Piwi subfamily of Arognaute proteins changed significantly and the Ping-Pong amplification cycle is impaired, suggesting that the Tudor proteins play a modulatory role in piRNA production (Vagin et al, 2009b).

### 1.2.5. 3' methylation of piRNAs (and *Drosophila* siRNAs) by Hen1 proteins

Whilst all small RNAs in *Arabidopsis thaliana* are methylated at the 3'- end, in the animal kingdom only piRNAs and *Drosophila* siRNAs are methylated on the 2'-hydroxyl group at the 3'- end. The enzymes responsible for such modification in

the animal kingdom have been identified to be the homologues of *Arabidopsis thaliana* methyltransferase HEN1 (Yu et al, 2005).

So far HEN1 homologues have been identified in animal (e.g. *Drosophila*, mouse, zebrafish) (Horwich et al, 2007) (Saito et al, 2007) (Ohara et al, 2007) (Kirino & Mourelatos, 2007a) (Kamminga et al, 2010), protozoa (e.g. *Tetrahymena thermophila*) (Kurth & Mochizuki, 2009) and bacteria (e.g. *Clostridium thermocellum* and *Anabaena variabilis*) (Chan et al, 2009) (Mui Chan et al, 2009) (Jain & Shuman, 2011). Apart from a conserved methyltransferase domain, these homologues have diverse sequences and domain structures, as shown in **Figure 8**. *Arabidopsis thaliana* methyltransferase HEN1 possesses two double-stranded RNA binding domains, a La Motif fold and a PPIase-Like domain at the N – terminus and a C- terminal methyltransferase domain, but there are no readily identifiable folds or domains predicted in the other HEN1 homologues other than the methyltransferase domain (Huang et al, 2009).

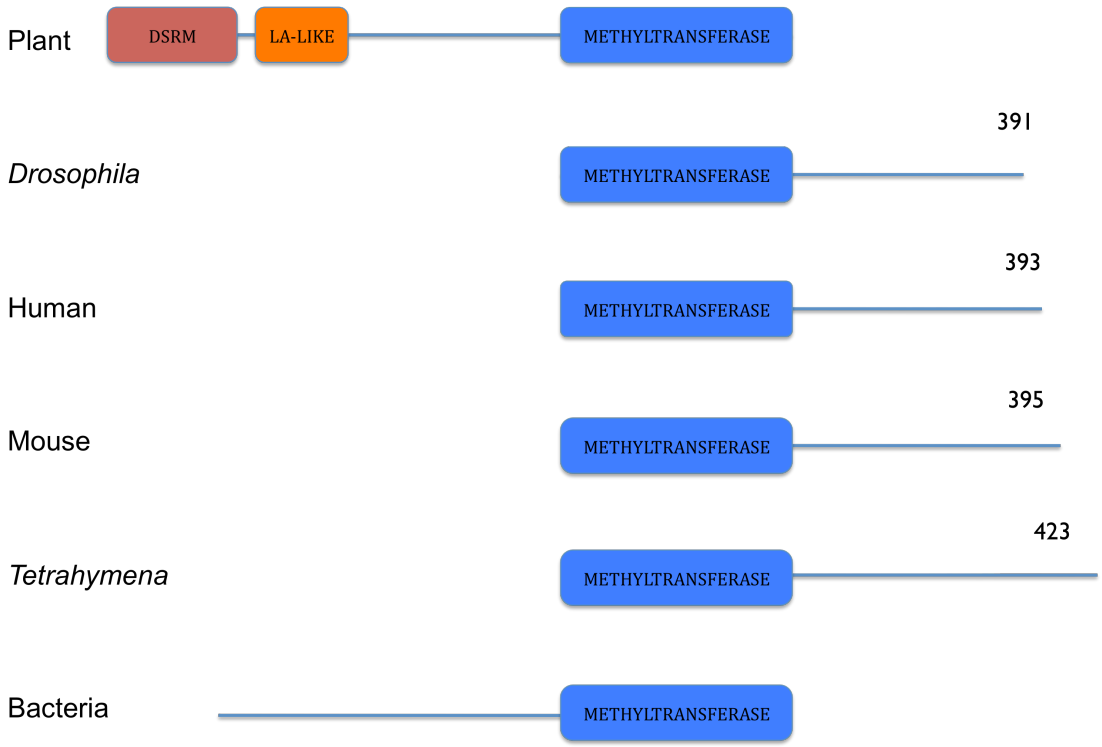



**Figure 8: Domain structure of subfamilies of Hen1 proteins**

Consistent with the lack of double stranded RNA binding domains, unlike the *Arabidopsis thaliana* methyltransferase HEN1 that acts on double stranded RNA, animal Hen1 homologues and *Tetrahymena thermophila* Hen1 methylate only single stranded RNAs (Yu et al, 2005) (Huang et al, 2009) (Vilkaitis et al, 2010) (Kirino & Mourelatos, 2007b) (Saito et al, 2007) (Horwich et al, 2007) (Kamminga et al, 2010) (Kurth & Mochizuki, 2009). Bacteria *Clostridium thermocellum* Hen1, on the other hand, showed a much broader substrate specificity as it demonstrated *in vitro* activities on both single stranded RNA, blunt-end or over-hang RNA duplex (Mui Chan et al, 2009) (Jain & Shuman, 2011).

In plant, methylation on the small RNAs at their 3' – ends prevents 3' uridylation, which acts as a signal for degradation (Li et al, 2005). In bacteria, Hen1 is thought to be closely involved with another RNA end-healing enzyme Pnkp in a RNA repair pathway. It is thought that by methylating the 3' terminal nucleotide of broken RNAs, Hen1 marks the broken ends of RNA at the repair junction and protects it from further damage by exonucleases activities (Chan et al, 2009) (Jain & Shuman, 2011) (Jain & Shuman, 2009). The protozoan *Tetrahymena thermophila* Hen1 is believed to be responsible for 2'-O-methylation on a selected class of small RNAs termed scnRNAs, expressed exclusively during sexual reproduction. The modification is thought to be responsible for scnRNA stabilisation, DNA elimination and subsequently survival of sexual progeny (Kurth & Mochizuki, 2009). In *vivo* studies in zebrafish showed that Hen1 was required for ovary development and fully efficient transposon silencing. It prevents uridylation-mediated exonuclease activity on 3' end of piRNA by adding a methyl group on the 2'-hydroxyl group at the 3' end of piRNA (Kamminga et al, 2010). In *Drosophila*, piRNA 3' – end methylation was shown to be required for normal piRNA function in *vivo* and the siRNA 3'- end methylation was thought to be the final step in assembly of the RISC – the effector complex in RNA silencing (Horwich et al, 2007) (Ameres et al, 2010). When supplemented with ovary lysate, the recombinant *Drosophila* Hen1 protein is much more active than it is *in vitro,* suggesting that the enzyme requires additional factor for full activation (Saito et al, 2007).

Little is known about the mechanism of piRNA 3'- end methylation in animals. piRNAs and *Drosophila* siRNAs are methylated by Hen1 when bound to Argonaute, after the dissociation of passenger strand. How Hen1 recognise the

substrate (Argonaute / small RNAs) and whether or not it is regulated is unknown. How does it avoid methylating other RNAs? The structure of any animal Hen1 is unknown.

## 1.3.The focus and the aim of the thesis: piRNA methylation

One of the fundamental questions is how the 3'- end of the piRNA is generated and recognised specifically by Piwi but not by the Ago subfamily of Argonaute proteins. Furthermore, the significance of the ubiquitous 2'-O-methylation modification at the 3'- end of the piRNAs that marks piRNA maturation is also yet to be elucidated.

Thus the aim of the research is to address these questions through:

1 – studying the structure, mechanism and regulation of the methyltransferase Hen1, responsible for the piRNA 2'-O-methylation at 3' end illustrated in **Figure 9**.

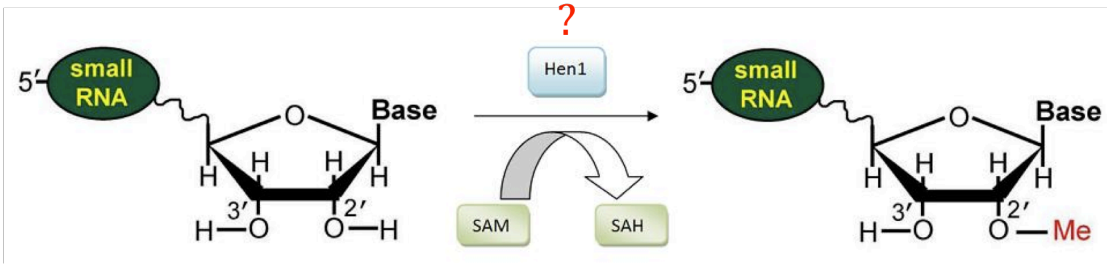



**Figure 9: 3' piRNA end methylation by methyltransferase Hen1 in the presence of co – factor S-Adenosyl Methionine (SAM)**

2 – Studying the structure and interaction of the methylated 3'- piRNA end with the PAZ domain of Piwi subfamily Argonaute proteins as illustrated in **Figure 10**.

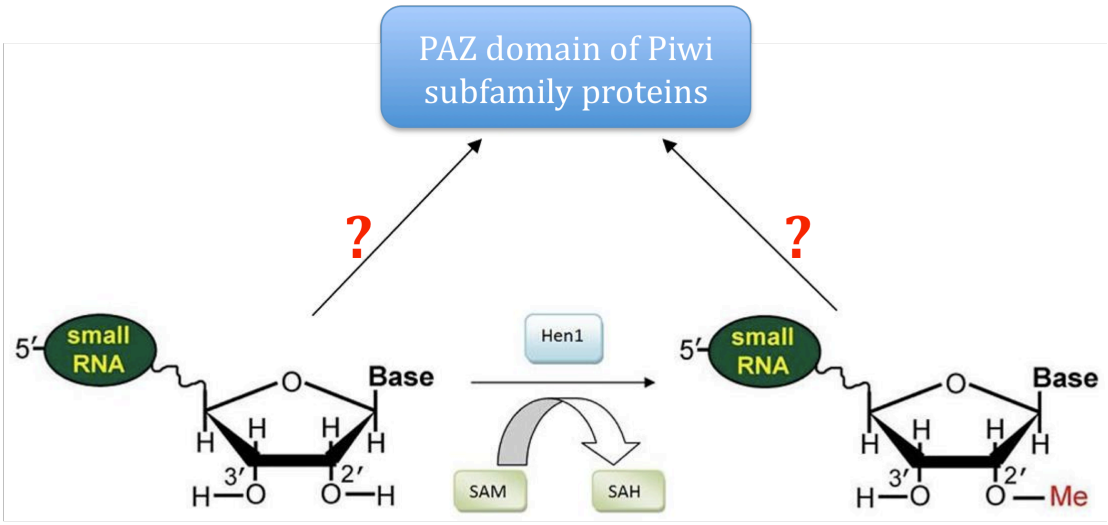



**Figure 10: Interactions between the 3'-end of piRNA, and the PAZ domain of Piwi subfamily Argonaute Proteins**

# 2. Methods and Materials

## 2.1.Introduction

The first section of this Chapter describes project specific methods and materials for Hen1 protein and Aubergine PAZ. The second section of this chapter describes the generic methods and materials.

## 2.2.Specific materials and methods for Hen1 and Aubergine PAZ

### 2.2.1. Cloning

#### 2.2.1.1. Hen1

The cDNAs of human Hen1, *Drosophila* Hen1 and zebrafish Hen1 were purchased from the MGC (mammalian Gene Collection), the DGC (*Drosophila* Gene Collection), and the ZGC (*Zebrafish* Gene Collection). The full length or truncated Hen1 DNA was amplified using primers listed in **Table 1**. Details of the expression vectors pTWO-E (made by Tony Oliver Institute of Cancer Research), pGEX-6P-1 (GE healthcare) and pET-30a (Novagen) are given in the appendix.

| Construct | Species | Residue Number | Expression vector† | Tag | Forward primer (5'-3') | Backward primer (5'-3') |
|---|---|---|---|---|---|---|
| hsHen1 | human | 1-393 | pTWO-E | His | GGAATTCCCATATGGAAGAAAATAATCTACAGTGC | GGAATTCCGCGGCCGCTCAAAACTCAAACTGTTCATC |
| drHen1 | zebrafish | 1-418 | pTWO-E | His | GGAATTCC CATATG ACCGCCACACCGTTCTCTCC | GGAATTCC GCGGCCGC TTATTCGTCTCCATAGCTACC |
| drHen1.2 | zebrafish | 1-418 | pGEX-6P-1 | GST | GGAATTCC GGATCC ATGACCGCCACACCGTTCTCTCC | GGAATTCC GCGGCCGC TTATTCGTCTCCATAGCTACC |
| dHen1 | *Drosophila* | 1-391 | pTWO-E | His | GGAATTCCCATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTATGATTCGGGGCCTTG |
| dHen1.2 | *Drosophila* | 1-391 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTATGATTCGGGGCCTTG |
| dHen1.3 | *Drosophila* | 1-391 | pET-30a | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC CTCGAG TGATTCGGGGCCTTGATCCC |
| dHen1.4 | *Drosophila* | 1-356 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCCGCGGCCGCTTATTCCTGATCGGAACTAAC |
| dHen1.5 | *Drosophila* | 1-356 | pET-30a | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC CTCGAG TTATTCCTGATCGGAACTAAC |
| dHen1.6 | *Drosophila* | 1-303 | pTWO-E | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCCGCGGCCGCTTAATTGGGTTTGCTAACCAACG |
| dHen1.7 | *Drosophila* | 1-303 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCCGCGGCCGCTTAATTGGGTTTGCTAACCAACG |
| dHen1.8 | *Drosophila* | 1-303 | pET-30a | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC CTCGAG TTAATTGGGTTTGCTAACCAACGG |
| dHen1.9 | *Drosophila* | 1-253 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTACATATCCTTGCGAACG |
| dHen1.10 | *Drosophila* | 1-239 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTACTCAATCTCGGAGGATTC |
| dHen1.11 | *Drosophila* | 1-356 | pTWO-E | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTATTCCTGATCGGAACTAAC |
| dHen1.12 | *Drosophila* | 1-368 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTAACGATCAGAAAACTTG |
| dHen1.13 | *Drosophila* | 1-368 | pTWO-E | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTAACGATCAGAAAACTTG |
| dHen1.14 | *Drosophila* | 1-364 | pGEX-6P-1 | GST | GGAATTCCGGATCCATGTTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTACTTGTAAGGATCAGACC |
| dHen1.15 | *Drosophila* | 1-364 | pTWO-E | His | GGAATTCC CATATG TTTTCGCACAAGTTTATTTGC | GGAATTCC GCGGCCGC TTACTTGTAAGGATCAGACC |

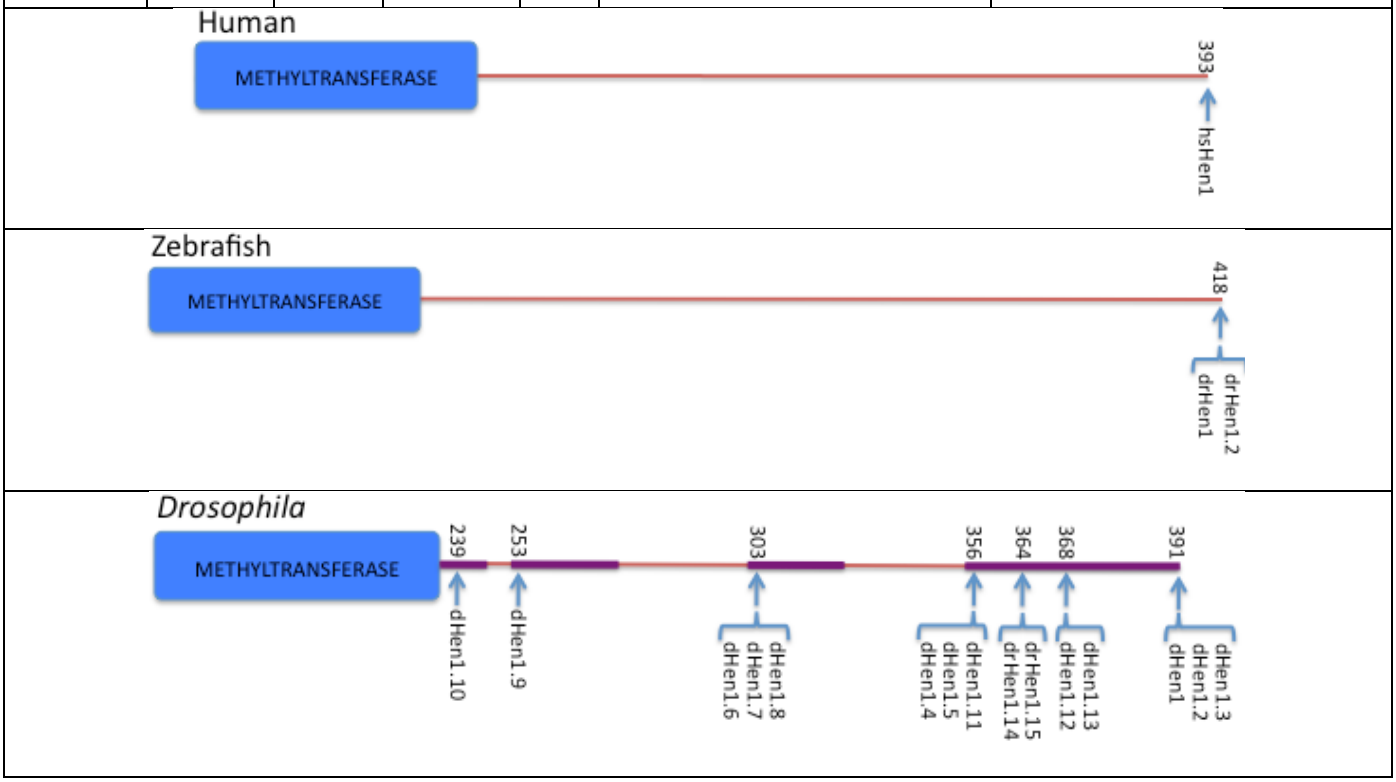



**Table 1: List of primers, constructs and schematic representations of constructs for Hen1 proteins expression**

#### 2.2.1.2. Cloning of PAZ constructs

The full length DNA encoding *Drosophila* Aubergine protein was a kind gift from Dr. J Parker. Subsequently, various constructs of Aubergine PAZ domains were cloned into pTWO-E vector. AubPAZ3$_{(279\text{-}398)}$ construct was also cloned into pOPINS vector (kind gift from Oxford Protein Production Facility (OPPF)) termed SumoPAZ using the In-Fusion® PCR cloning kit (from Clontech) following the manufacturer's protocol. Thus SumoPAZ and AubPAZ3$_{(279\text{-}398)}$ encodes for the same protein fragment with different expression tags. **Table 2** lists the primers for all constructs.

| Construct | Species | Residue Number | Expression vector† | Tag | Forward primer (5'-3') | Backward primer (5'-3') |
|---|---|---|---|---|---|---|
| AubPAZ1 | *Drosophila* | 274-398 | pTWO-E | His | GGAATTCC CATATG CACAAGGTTATGCGAACTGAG | GGAATTCC GCGGCCGC TTAGTCAGCGCGCATAGCGTC |
| AubPAZ2 | *Drosophila* | 274-412 | pTWO-E | His | GGAATTCC CATATG CACAAGGTTATGCGAACTGAG | GGAATTCC GCGGCCGC TTAATTCAGCCTGGTATGTTC |
| AubPAZ3 | *Drosophila* | 279-398 | pTWO-E | His | GGAATTCCCATATGACTGAGACCTTGTACAATATTTTG | GGAATTCCGCGGCCGCTTAGTCAGCGCGCATAGCGTC |
| AubPAZ4 | *Drosophila* | 279-412 | pTWO-E | His | GGAATTCCCATATGACTGAGACCTTGTACAATATTTTG | GGAATTCC GCGGCCGC TTAATTCAGCCTGGTATGTTC |
| SumoPAZ | *Drosophila* | 279-398 | pOPINS | His-Sumo | GCGAACAGATCGGTGGT GGGCCCCATATGACTGAGA | ATGGTCTAGAAAGCT TTAGTCAGCGCGCATAGC |

**Table 2: List of primers and constructs for Aubergine PAZ domain proteins**

#### 2.2.1.3. Site – Directed Mutagenesis

AubPAZ3$_{(279\text{-}398)}$ and SumoPAZ mutants were generated by PCR method using QuikChange II site-directed Mutagenesis Kit (from Aligent) following the manufacturer's protocol. **Table 3** lists the primers for all mutants.

| Construct | Mutation | Expression *† | Tag | Sense primer (5'-3') | Antisense primer (5'-3') |
|---|---|---|---|---|---|
| PAZ F55A | UUU-GCU | pTWO-E | His | CAATCGACGCCATTGTGCAAAGCTAAAACCAATGACGGTGAG | CTCACCGTCATTGGTTTTAGCTTTGCACAATGGCGTCGATTG |
| PAZ F55AY67A | UAC-GCC | pTWO-E | His | CGGTGAGATTTCGTACGTGGATGCCTATAAGAAGCGATACAAC | GTTGTATCGCTTCTTATAGGCATCCACGTACGAAATCTCACCG |
| PAZ F55AY67AY72A | UAC-GCC | pTWO-E | His | CGTGGATGCCTATAAGAAGCGAGCCAACATAATCATTCGCGACTTG | CAAGTCGCGAATGATTATGTTGGCTCGCTTCTTATAGGCATCCACG |
| SumoPAZMut1 (Y35F) | TAT-TTT | pOPINS | His-Sumo | GGGTATGGTAATACTAACGGATTTTAATAACAAAACCTATCGCAT | CAATGCGATAGGTTTTGTTATTAAAATCCGTTAGTATTACCATACCC |
| SumoPAZMut2 (N36A) | AAT-GCT | pOPINS | His-Sumo | CAGTAATGGGTATGGTAATACTAACGGATTATGCTAACAAAACCTT | ATGCGATAGGTTTTGTTAGCATAATCCGTTAGTATTACCATACCCATT |
| SumoPAZMut3 (N36C) | AAT-TGT | pOPINS | His-Sumo | CAGTAATGGGTATGGTAATACTAACGGATTATTGTAACAAAACCTT | ATGCGATAGGTTTTGTTACAATAATCCGTTAGTATTACCATACCCATT |
| SumoPAZMut4 (N36D) | AAT-GAT | pOPINS | His-Sumo | GGTATGGTAATACTAACGGATTATGATAACAAAACCTATCGCATT | GTCAATGCGATAGGTTTTGTTATCATAATCCGTTAGTATTACCATACC |
| SumoPAZMut5 (Y64I) | TAC-ATC | pOPINS | His-Sumo | CCAATGACGGTGAGATTTCGATCGTGGATTACTATAAGAAGC | GCTTCTTATAGTAATCCACGATCGAAATCTCACCGTCATTGG |
| SumoPAZMut6 (Y64V) | TAC-GTC | pOPINS | His-Sumo | CCAATGACGGTGAGATTTCGGTCGTGGATTACTATAAGAAGC | GCTTCTTATAGTAATCCACGACCGAAATCTCACCGTCATTGG |
| SumoPAZMut7 (Y67F) | TAC-TTC | pOPINS | His-Sumo | GTGAGATTTCGTACGTGGATTTCTATAAGAAGCGATACAACAT | ATGTTGTATCGCTTCTTATAGAAATCCACGTACGAAATCTCAC |
| SumoPAZMut8 (Y72F) | TAC-TTC | pOPINS | His-Sumo | GTGGATTACTATAAGAAGCGATTCAACATAATCATTCGCGACTTG | CAAGTCGCGAATGATTATGTTGAATCGCTTCTTATAGTAATCCAC |

**Table 3: List of constructs and primers for the Aubergine PAZ domain protein mutants**

## 2.2.2. Protein expression and purification

### 2.2.2.1. General

Transformed *E. coli* cells were grown in typically 12 litres of Luria Broth (LB) and protein expression was induced by adding Isopropyl-β-s-thiogalactopyranoside (IPTG). Cells were harvested by centrifuging at 5000 RPM for 20 minutes and then re-suspended in the lysis buffer before flash freezing in liquid nitrogen for storage at -80°C. Cell lysis was achieved by sonication and cell debris was removed by centrifugation. Protein purification was performed on

AKTA FPLC chromatographic systems controlled by the UNICORN software (GE Healthcare) in the cold cabinet at 4°C and the purified protein was stored at -80°C after flash freezing.

RNase free samples are typically obtained at the final gel filtration stage. The final gel filtration buffer was made using RNase free chemical, MilliQ water, RNase free pipette tips and buffer filters. The protein sample was stored in RNase free eppendorfs (Dharmacon).

#### 2.2.2.2. Hen1 proteins

Full length and truncated constructs of Hen1 protein were expressed in ArcticExpress (DE3) competent cells (Stratagene). Transformed *E. coli* cells were grown in 12L of LB at 37°C till $OD_{600}$ = 0.9 and the protein expression was induced by adding 0.1 mM IPTG and cells were grown at 16°C overnight. Cells were harvested by centrifugation at 5000 RPM for 20 minutes at 4°C and the cell pellets were resuspended in loading buffer (50 mM Tris pH7.4, 150 mM NaCl) supplemented with 2 mM $MgCl_2$, 3 Complete Protease Inhibitor cocktail tablets (Roche, including Serine and Cysteine Protease inhibitors), 1 mg/ml lysozyme and 0.3 mg DNaseI. The cells were lysed by sonication and the cell debris was removed by centrifugation at 20,000 RPM for 1 hour at 4°C. The soluble fraction was loaded onto a gravity glutathione column and the GST –fusion protein was eluted with elution buffer (50 mM Tris pH7.4, 150 mM NaCl, 2 mM Dithiothreitol (DTT) and 20 mM glutathione) after washing with 200 ml loading buffer. The GST – tag was cleaved by 3C protease overnight at 4°C and the cleaved product was collected by running a second glutathione column to remove GST tag and uncleaved fusion protein. The eluted protein was subsequently buffer exchanged

into the monoQ buffer A (50 mM Tris pH8.5, 50 mM NaCl and 3 mM DTT) using a HiPrep™ 26/10 desalting column (GE Healthcare) and then further purified through monoQ 5/50GL column (GE Healthcare) using a salt gradient of 0 – 100% monoQ buffer B (50 mM Tris pH8.5, 1 M NaCl and 3 mM DTT) over a volume of 60 ml. Finally the purified protein was eluted in 'sample buffer' (10 mM Tris–HCl pH7.4, 150 mM NaCl, 5 mM DTT and 2 mM EDTA) on size exclusion superdex200 gel filtration column.

#### 2.2.2.3. PAZ

Both AubPAZ3$_{(279\text{-}398)}$ proteins expressed in pTWO-E vector and SumoPAZ protein expressed in pOPINS vector employed a similar large scale expression and purification scheme. Transformed *E. coli* cells (BL21 (DE3) pLysS competent cells (Novagen) were grown in 6 Litres of LB at 37°C till $OD_{600}$ = 0.6 and the protein expression was induced by adding 0.1 mM IPTG and cells were grown at 18°C overnight. Cells were harvested by centrifugation at 5000 RPM for 20 minutes at 4°C and the cell pellets were resuspended in Loading Buffer (50 mM Hepes pH7.6, 150 mM NaCl) supplemented with 2 mM $MgCl_2$, 2 Complete Protease Inhibitor cocktail tablets (Roche, Serine and Cysteine protease inhibitors), 1 mg/ml lysozyme and 0.3 mg DNaseI. The cells were lysed by sonication and the cell debris was removed by centrifugation at 20,000 RPM for 1 hour at 4°C. The supernatant supplemented with 500 mM NaCl and 30 mM imidazole was then loaded onto a Ni-NTA column, washed with 300 ml Ni Buffer A (50 mM Hepes pH7.6, 500 mM NaCl) and 10% Ni Buffer B (50 mM Hepes pH7.6, 500 mM NaCl, 300 mM Imidazole) and eluted using an imidazole gradient of 10 – 100% Ni Buffer B over a volume of 100 ml. To each fraction 2 mM DTT

and 1 mM EDTA was added. Pooled fractions of AubPAZ3$_{(279-398)}$ proteins were cleaved by 3C protease overnight at 4°C followed by putting the sample through a gravity glutathione column to remove 3C protease. Pooled fractions of SumoPAZ proteins were cleaved by Sumo protease overnight at 4°C. The cleaved product was then desalted back into Ni Buffer A using the desalting column and purified further through a second Ni-NTA column eluting with an imidazole gradient of 0-100% Ni Buffer B over a volume of 150 ml. Finally the purified protein was eluated in 'sample buffer' (5 mM Hepes pH7.5 150 mM NaCl, 5 mM DTT and 2 mM $MgCl_2$) on a size exclusion superdex75 gel filtration column.

### 2.2.3. Biochemical and biophysical assays

#### 2.2.3.1. Hen1 In *vitro* methylation assays

##### 2.2.3.1.1. With detection by autoradiography

Purified dHen1 protein (2 μg) in a 50 μl reaction mixture were incubated with single-stranded RNAs (500 pmol) at either 26°C or 37°C in a buffer containing 50 mM Tris-HCl (pH8.0), 100 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 5% Glycerol, 0.5 μl RNasin (Promega) and 0.15 μCi S-adenosyl-L-(methyl-$^{14}$C) methionine (SAM) (GE Healthcare). After a 2 hour incubation, the reaction was stopped by adding 1 μl of proteinase K and heated at 65°C for 20 minutes followed by 2 minutes on ice. The RNAs were isolated by phenol extraction and loaded to 20% 7M Urea Containing denaturing polyacrylamide gels. The RNA was visualised by TBO (Toluidine Blue O) staining and the gel was treated with the autoradiography enhancer (En$^3$hance from Perkin Elmer) following the manufacturer's instructions and exposed to film at -80°C for 3 days. The reaction resulted in the transfer of the $^{14}$C-methyl group from SAM (S-adenosyl-

methionine) to the single-stranded RNAs, which become $^{14}C$-labeled and could be visualised on the film upon developing.

#### 2.2.3.1.2. With detection by liquid scintillation counting

The initial part of the methylation assays was carried out in the same way as described above. After proteinase K treatment, triplicate samples were spotted onto 2.3 cm 3MM filters (Whatman), washed two times with 0.05M sodium phosphate buffer, then two times with water, two times with ethanol and once with acetone followed by complete drying. The $^{14}C$ radioactivity was measured by liquid scintillation counting in 4 ml of scintillation cocktail (Ultima Gold MV from Perkin Elmer) using a liquid scintillation spectrometer.

### 2.2.3.2. Analytical gel filtration

Analytical gel filtration was performed on a Superdex200 10/300 column and Superdex75 10/300 column with sample loading volume of 100 µl on AKTA FPLC chromatographic systems controlled by the UNICORN software (GE Healthcare) in the cold cabinet at 4°C. Five known protein standards: Thyroglobulin (670 kDa), BSA dimer (132 kDa), BSA monomer (66 kDa), Ovalbumin (44 kDa) and Equine Myoglobin (17 kDa) were used for calibration of the Superdex200 column and Thyroglobulin (670 kDa), Albumin (44 kDa), Carbonic Anhydrase (29 kDa), Cytochrome c (12.4 kDa) and Aprotinin (6.5 kDa) were used for calibration of the Superdex75 column.

#### 2.2.3.3. Analytical Ultracentrifugation

Sedimentation velocity experiments and analysis were conducted at 20°C in a Beckman Optima XL-A analytical ultracentrifuge in house at the department of Biochemistry assisted by Dr David Staunton using SEDFIT analysis software. The buffer in the experiment contained 5 mM Hepes pH7.5 150 mM NaCl, 5 mM DTT and 2 mM $MgCl_2$. 120 μl of 40 μM protein sample was spun at 25,000 RPM and 30,000 RPM and the 280 nm absorbance reading data was collected.

#### 2.2.3.4. Isothermal titration calorimetry (ITC)

ITC was conducted on ITC200 MicroCalorimeter at 25°C. Protein samples were dialysed into sample buffer containing 50 mM Tris pH7.5, 150 mM KCl, 5% (v/v) Glycerol, 10 mM $MgCl_2$, 1 mM DTT at 4°C overnight prior to ITC experiment. The RNA samples were directly diluted from concentrated stock into buffer to match the exact composition of the protein sample buffer to ensure that the only heat measured is due to the solute. ITC analysis was performed by first injecting 20 x 2.5 μl 50 μM RNA into 250 μl sample buffer in the instrumental cell; then repeating the experiment by replacing the sample buffer with 25 μl 5 μM $AubPAZ3_{(279\text{-}398)}$ protein. The data were analysed using MicroCal Origin software.

#### 2.2.3.5. Electrophoretic mobility shift assay (EMSA.)

20 μl 250 nM RNA oligo was end-labelled with [γ- $^{32}P$] ATP using T4 PNK (NEB) by incubating the mixture containing RNA oligo, 10x PNK buffer (NEB) and T4 PNK at 37°C for 1 hour. The labelled oligo was purified using a G-25 spin column (Amersham) equilibrated with 10 mM Hepes pH7.5.

The binding reaction was performed on ice for 30 minutes in 50 mM Hepes pH7.5, 250 mM NaCl, 5 mM $MgCl_2$, 2 mM DTT, 50 ng/μl poly dI-dC with 0.5 fmol/μl labelled RNA oligo. Each binding reaction contains a concentration of protein from a serial protein dilution. 10% Glycerol was added to each reaction mixture prior to loading the samples on 10% non-denaturing polyacrylamide gels in 1 x TBE buffer. The gel was fixed in 7% acetic acid for 15 minutes and dried before exposing for phosphorimaging.

#### 2.2.3.6. Fluorescence Anisotropy

Fluorescence anisotropy was performed on Pherostar microplate reader (BMG labtech). The binding reaction was carried out in buffer containing 50 mM Hepes pH7.5, 250 mM NaCl, 5 mM $MgCl_2$ and 2 mM DTT. 10 nM fluorescently labelled RNA oligos were incubated with $AubPAZ3_{(279\text{-}398)}$ protein with a concentration range of 0 – 160 μM. 20 μl of each reaction mix was pipetted into a low protein binding 384-well dark plate and the fluorescence anisotropy reading was measured at 20°C using the fluorescein polarisation module with excitation wavelength of 485 nm and emission wavelength at 520 nm.

### 2.2.4. RNA oligonucleotides

#### 2.2.4.1. Hen1

All RNA oligonucleotides are commercially synthesized and obtained from Dharmacon.

The RNA oligonucleotides used in crystallisation and Hen1 activity assays are summarised in **Table 4**.

| Name: | Sequence (5’ – 3’) |
|---|---|
| Let-7b 22nt | UGAGG UAGUA GGUUG UGUGG UU |
| Let-7b 14nt | UA GGUUG UGUGG UU |
| Let-7b 10nt | UUG UGUGG UU |
| Let-7b 5nt | UGG UU |
| 3mer | CAC |
| 2mer | AC |

**Table 4: List of single stranded RNA sequences used with Hen1 proteins**

2.2.4.2. PAZ

The RNAs used in ITC, band shift assay, fluorescence anisotropy experiment and crystallisation trials are listed in **Table 5**.

| Name: | Length | Modification | Sequence (5’ – 3’) |
|---|---|---|---|
| RNA1 | 7 | 2’-O-methylation at 3’-end | AACCUAC-Met |
| RNA2 | 7 | None | AACCUAC |
| RNA3 | 7 | 5’-Fluorescein and 3’- end 2’-O-methylation | Fl-AACCUAC-Met |
| RNA4 | 7 | 5’-Fluorescein | Fl-AACCUAC |
| RNA5 | 7 | 5’-Fluorescein and 3’- end 2’-O-methylation | Fl-ACAAGCA-Met |
| RNA6 | 7 | 5’-Fluorescein | Fl-ACAAGCA |

**Table 5: List of sequences of single stranded RNAs used with Aubergine PAZ domain proteins. m - methyl on the 2'-OH group at the 3'-terminal nucleotide; Fl - Fluorescein on the first nucleotide at the 5'- end**

### 2.2.5. Crystallization trials

2.2.5.1. 96-well robotic screens

Mosquito and Tecan Robots were used to set up sitting drops in 96-well trays with typical drop volume of 200nl (1:1 sample: precipitant ratio) and precipitant volume of 100μl at 4°C and 20°C. The standard screens including PACT premier

(Molecular Dimensions), JCSG+ (Molecular Dimensions), Molecular Dimensions Structure Screens 1+2 (Molecular Dimensions), Index (Hampton Research), PEG/Ion 1+2 (Hampton Research) and Wizard 1+2 were used (Emerald Biosystems).

#### 2.2.5.2. 24-well screens

24-well hanging drop screens were set up by hand with a typical drop volume of 2 μl (1:1 sample: precipitant ratio) and precipitant volume of 600 μl at 4°C and 20°C. The trays and cover slips (CrystalClene™) used were both RNase free (Molecular Dimensions). For each precipitant condition, up to three drops of apo or complexed sample could be set up on the same cover slip. Further optimisation of the crystals were set up by varying precipitant concentration, pH and the addition of additives.

## 2.3.Generic materials and methods

### 2.3.1. Transformation

50 μl competent cells were carefully thawed on ice and mixed with 50 ng of plasmid DNA for 25 minutes. After 30 seconds of heat shock in a 42°C water bath the cells were transferred immediately back onto ice for another 2 minutes. Subsequently 0.9 ml of sterile LB was added to the cells and incubated for 1 hour at 37°C with vigorous shaking. Finally the cells were spun down at 8000 RPM, resuspended in 100 μl of sterile LB and plated out onto an LB agar plate containing the appropriate antibiotics for overnight growth at 37°C.

### 2.3.2. Polymerase Chain Reaction -PCR

PCR reaction was carried out in a 50 μl reaction mix containing 50 ng of template DNA, 1 μl 100 μM each of forward and reverse primers, 1 μl 200 μM dNTPs, 5 μl 10x thermal reaction buffer and 1 μ Vent DNA polymerase (NEB). A typical PCR program started with initial denaturing step at 95°C for 3 minutes, then 24 cycles of denaturation (at 94°C for 15 seconds), annealing (at 55°C for 30 seconds) and extension (72°C at a rate of 1 kb/min). The reaction mix was cooled to 4°C after a final extension of 10 minutes at 72°C. The PCR product was checked by running 1% agarose gel electrophoresis in TAE buffer (40 mM Tris pH8.5, 20 mM acetate, 1 mM EDTA) and was purified using the Qiaquick PCR clean-up kit (Qiagen).

### 2.3.3. Restriction enzyme digest and ligation

The desired coding sequences of proteins of interest were amplified by PCR using primers containing restriction sites suitable with the vector multiple cloning site. 200 μl of reaction mix containing 10 μg of PCR product, 50 μg of vector, 1 μl of calf alkaline phosphatase (NEB), 5 μl of each restriction enzyme (NEB) supplemented with 20 μl of 10× reaction buffer and 2 μl of 100× BSA were incubated at 37°C for 1 hour. The reaction mix was then subject to Proteinase K treatment to remove active enzymes by adding 20 μl Proteinase K (Sigma) and 30 μl of 10% Proteinase K digestion buffer (10 mM Tris pH 7.8, 5 mM EDTA, 0.5% SDS,) at 65°C for 20 minutes. The digestion products were purified by running 1% agarose gel electrophoresis and gel extraction using MinElute Gel Extraction Kit (Qiagen). Ligation of insert and linearised vector was carried out in a 20 μl reaction mix containing 100 ng of vector, 3× molar excess of insert

using the formula shown in **Equation 1**, 2 μl of 10× T4 ligation buffer and 1 μl of T4 DNA ligase (NEB) at 20°C for 1 hour.

$$\text{ng of insert to use} = \text{3 (Molar ratio of insert/ vector)} \times \frac{\text{Size of Insert}}{\text{Size of Vector}} \times \text{100 ng (amount of vector to use)}$$

**Equation 1: Formula for quantifying the insert used for ligation**

Ligated products of expression vectors and the desired inserts were transformed into DH5α competent cells (Invitrogen) and incubated overnight. A single colony was picked and incubated in 10 ml LB supplemented with appropriate antibiotics shaking overnight at 37°C and the cell pellet was harvested by centrifugation. The DNA was then extracted using a Miniprep kit (Qiagen)

### 2.3.4. Western Blotting

Western blotting used during expression trials followed the standard protocol described in Molecular Cloning: A Laboratory Manual (Sambrook and Russell, 2001). The reagents used are listed in **Table 6**:

| Reagent | Description |
|---|---|
| Primary anti-His tag antibody | His-tag monoclonal antibody (Novagen Catalogue No. 70796-4 ) – mouse monoclonal antibody $IgG_1$ |
| Primary anti-GST tag antibody | GST-tag monoclonal antibody (Novagen Catalogue No. 71097-3) – mouse monoclonal antibody $IgG_1$ |
| Secondary antibody | Anti-Mouse IgG- Alkaline Phosphatase-conjugated antibodies (Promega) |
| Alkaline Phosphatase substrate | SIGMA FAST BCIP/NBT (5-Bromo-4-chloro-3-indolyl phosphate/Nitro blue tetrazolium) tablets – precipitating substrate for the detection of alkaline phosphatase activity. |

**Table 6: List of reagents used in Western blot**

### 2.3.5. SDS-PAGE

Protein expression and purity were assessed by running samples on SDS-PAGE gels. Equal volume of sample and 2x SDS -loading buffer were mixed and heated at 100°C for 2 minutes immediately prior to loading onto the NubPAGE 4-12% Bis-Tris gel (Invitrogen). Gels were run in NuPAGE Mes buffer (Invitrogen) typically at 200V for 35 minutes and protein was stained and fixed in solution containing 10% (v/v) acetic acid, 10% (v/v) ethanol, 0.05% (w/v) Coomassie Brilliant Blue for 5 minutes before being destained in solution containing 10% (v/v) acetic acid, 10% (v/v) ethanol.

### 2.3.6. **Websites for disorder prediction and secondary structure prediction are listed below**:

- RONN (Regional Order Neural Network) Disorder Prediction:

  http://www.strubi.ox.ac.uk/RONN

- Jpred secondary structure prediction program:

  http://www.compbio.dundee.ac.uk/www-jpred/

- Phyre secondary structure prediction program:

  http://www.sbg.bio.ic.ac.uk/phyre/

- DINAMelt Server for prediction of Melting profiles for nucleic acids

  http://www.bioinfo.rpi.edu/applications/hybrid/twostate.php

- ProtParam tool for calculating protein molecular weight, theoretical pI value, amino acid composition and extinction coefficient:

  http://www.expasy.org/protparam/

### 2.3.7. Protein and nucleic acid concentration determination.

#### 2.3.7.1. Protein concentration determination

Protein concentration was determined by measuring its absorbance at 280 nm using Nanodrop -1000 spectrophotometer. The theoretical extinction coefficient of protein at 280 nm was calculated using the ProtParam tool from Expasy online services.

The ratio of the absorbance reading at 260 nm/280 nm was an indicator for nucleic acids contamination – typically if the 260 nm/280 nm ratio was higher than 0.6 then it implied that the protein was associated with nucleic acids. This is important when purifying nucleic acid binding proteins.

Protein concentration was also determined by Bradford Assay using Coomassie protein assay reagent (Biorad) and measuring its absorbance at 595 nm when

estimating total protein concentration after the first step of protein purification and the Nanodrop spectrum was interfered by the presence of imidazole or glutathione. Standard Bovine serum albumin (BSA) proteins of known concentrations were used for calibration.

### 2.3.8. Nucleic acid concentration determination

Nucleic acid concentration was determined by measuring its absorbance at 260 nm using the Nanodrop -1000 spectrophotometer. The theoretical extinction coefficient of the RNA was provided by the commercial manufacturer (Dharmacon).

### 2.3.9. Small scale expression and solubility testing

The construct of DNA was transformed into appropriate expression strain competent cells (e.g. BL21 (DE3)) overnight according to the manufacturer's protocol. A single colony was picked and inoculated into 50 ml of LB culture supplemented with antibiotics and the protein expression was induced by adding 0.1 mM, 0.5 mM or 1.0 mM IPTG when cell growth was at the optimal density ($OD_{600}$ = 0.6 or 0.9). The induction conditions were tested at 18°C overnight, 25°C overnight and 37°C for 3 hours. The cell pellet was harvested and resuspended in lysis buffer (50 mM Tris pH7.4, 150 mM NaCl, 2 mM $MgCl_2$, Complete Protease Inhibitor cocktail tablets (Roche), 1 mg/ml lysozyme and 0.3 mg DNaseI.) Soluble fraction was obtained through cell lysis by sonication for 1 minute at an output power of 6 Watts and cell debris removal by centrifugation

at 13000 RPM for 30 minutes. Western blotting was performed on the total and soluble fractions of the cell lysate.

### 2.3.10. 3C-protease and Sumo Protease

3C-protease (PreScission protease, a 46 kDa fusion protein of GST and human rhinovirus 3C protease) was expressed and purified using a GST-affinity column followed by desalting into storage buffer. Purification was carried out on AKTA FPLC. Sumo Protease was a kind gift from Dr James Parker.

### 2.3.11. Limited proteolysis

0.5 mg/ml protein sample was incubated in reaction buffer (90 mM Tris pH8.5, 4 mM $CaCl_2$, 4 mM DTT) with 3× dilution series of either Trypsin (Promega) or Subtilisin (Sigma) at room temperature for 1 hour. The reactions were stopped by adding 2× SDS loading buffer (0.5M Tris-HCl pH6.8, 4.4% (w/v) SDS, 20% (v/v) glycerol, 2% (v/v) 2-mercaptoethanol, and bromophenol blue) and heating at 95°C for 2 minutes. The samples were then run on SDS-PAGE gel.

### 2.3.12. N-terminal sequencing

Limited proteolysis revealed the appropriate amount of Trypsin or Subtilisin for the generation of a detectable and stable digested fragment. N-terminal sequencing sample was prepared by incubating 0.5 mg/ml full length protein with the appropriate amount of Trypsin or Subtilisin at room temperature for 1 hour in reaction buffer. The reaction was stopped by adding 2× SDS loading buffer and heating at 95°C for 2 minutes. N-terminal sequencing (both gel and sequece analysis) was then performed in house in the Department of Biochemistry by Dr. Antony Willis.

### 2.3.13. Mass spectrometry

ESI-MS (Electrospray Ionisation Mass spectrometry) was performed in house in the department of Biochemistry by Dr. David Staunton. The sample was prepared using C4 ZipTip® (Millipore) to bind and eluate protein into 80% acetonitrile and 1% TFA solution for analysis.

# 3. Results and Discussion

## 3.1. Hen1 Expression, Purification and Crystallisation Trials

### 3.1.1. Introduction

Hen1 protein is the enzyme responsible for methylating the 3'-end of the piRNAs and the *Drosophila* siRNAs. The structure of animal Hen1 is unknown and the mechanism of methylation is unclear. Sequence alignments of *Drosophila*, human, mouse, *Tetrahymena* and *Arabidopsis* Hen1 proteins indicate that the methyltransferase domain for animal Hen1 takes up approximately the first half of the amino acid sequence and that the C-terminal region of the animal Hen1 proteins do not possess any known protein folds and are diverse in sequence amongst the animal Hen1 protein families as shown in **Figure 11**.

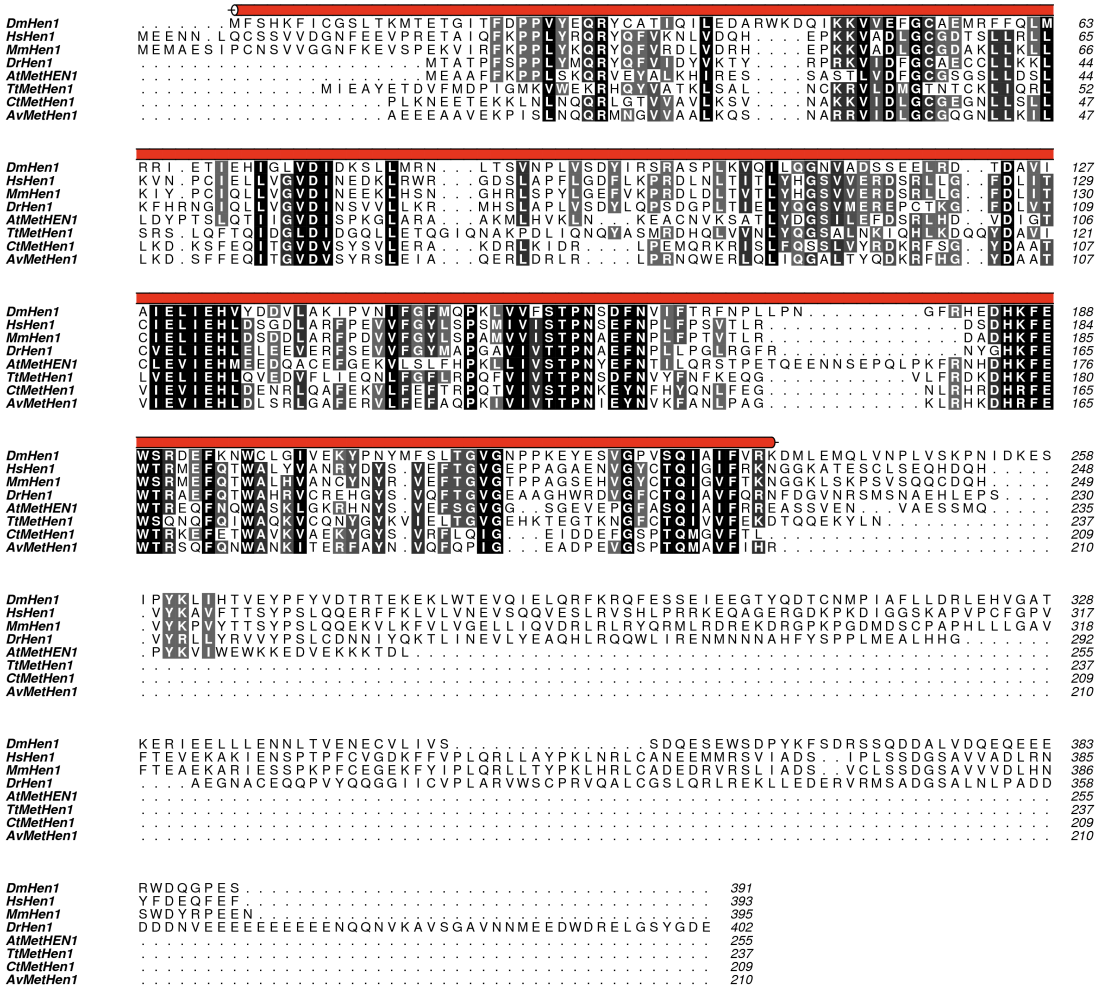



**Figure 11: Sequence alignment of full length animal Hen1 proteins (Dm - *Drosophila*, Hs - Human, Mm - Mouse, Dr - Zebrafish) with methyltransferase domain from plant (At Met), *Tetrahymena thermophila* (ttMet), bacteria *Clostridium thermocellum (Ct*Met) and *Anabaena variabilis* (*Av*Met). The red bar across the top indicates the predicted methyltransferase domain for animal Hen1 protein.**

The *Drosophila* Hen1 methyltransferase domain lies approximately between residues 1 to 237; for human Hen1 the methyltransferase domain lies between residues 1 to 232 and for mouse Hen1 methyltransferase domain lies approximately between residues 1 to 233.

The aim of the work detailed in this section was to produce sufficient quantity of biologically active and soluble animal Hen1 protein for structural and functional studies. This required making suitable Hen1 constructs for expression followed by establishing purification strategies to obtain homogenous Hen1 protein to

maximise the chance of crystallisation. The ultimate goal was, by determining the structure of an animal methyltransferase Hen1, to elaborate the function of piRNA/siRNA methylation, and gain mechanistic insights into piRNA biogenesis in animal.

### 3.1.2. Strategy

Flexile and unstructured regions in the protein as well as the heterogeneity of the protein sample could hinder crystallisation; and normally a significant amount of protein is required for sampling the crystallisation conditions. Therefore in order for the structural study of Hen1 protein, a suitable and stable fragment that retains the enzymatic activity and also expresses reasonably well needs to be identified. Usually the product of limited proteolysis corresponds to a stable construct that is more likely to crystallise.

### 3.1.3. Initial expression studies

#### 3.1.3.1. Choice of species

In initial expression trial, full-length human, zebrafish and *Drosophila* Hen1 were tested and it showed that *Drosophila* Hen1 (dHen1) expressed more solubly and stably in E. *coli* compared to human and zebrafish Hen1 with 0.1 mM IPTG and 16°C expression overnight. The expected molecular weight of the recombinant full-length *Drosophila*, human and zebrafish Hen1 were 47.5kDa, 46.5kDa and 48.1kDa respectively. **Figure 12** shows the soluble fraction of full length *Drosophila*, human and zebrafish Hen1 expression on both SDS-PAGE gel with Coomassie staining and western blot. The red boxes with expected molecular weights are shown for the three species of Hen1 protein bands. Therefore dHen1 was pursued.

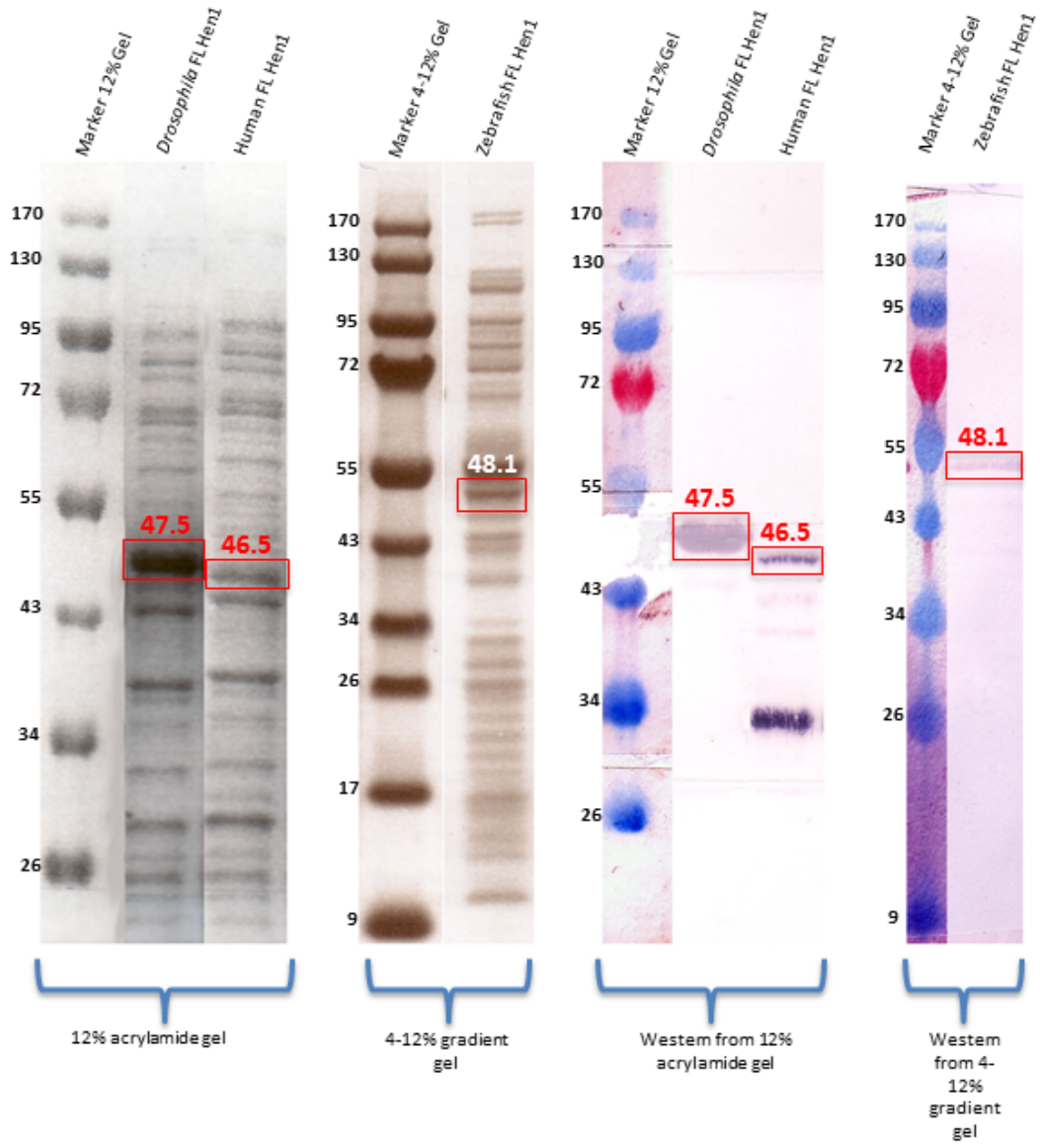



**Figure 12: Small scale solubility testing for Hen1 proteins across human, *Drosophila* and zebrafish. The left five panels show the SDS-PAGE staining of the soluble fractions and the right hand five panels show the corresponding western blot of the soluble fractions. The red box with expected molecular weight indicate the Hen1 proteins**

### 3.1.3.2. Purification of full length dHen1

Both N-terminal His-tagged (dHen1.1$_{(1\text{-}391)}$) and GST-tagged (dHen1.2$_{(1\text{-}391)}$) full-length dHen1 failed to bind to the nickel and GST resin respectively. Protein remained mostly in the flow through fractions, indicating that the tags in both proteins were inaccessible, which suggested the protein was aggregated.

### 3.1.4. Truncated constructs for dHen1

C- terminal region of *Drosophila* Hen1 does not have any known structural folds, thus this region could be disordered and potentially cause protein aggregation. The fragment of full-length dHen1 protein with C- terminal truncation hence may be less aggregated than the full length dHen1 protein.

#### 3.1.4.1. RONN disorder prediction

Flexible regions in the protein inhibit crystallisation. The RONN program predicted the probability of regions of protein with flexible loops. A schematic representation of dHen1 domain structure based on RONN disorder prediction program is shown in **Figure 13,** where the purple parts are predicted disordered regions. It indicated that the C-terminal tail of residue 355 to 391 is highly disordered. In addition, there are another three regions showing some level of disorder, which are comprised of residues 182-187 in the catalytic domain, 247-256 and 291- 306.

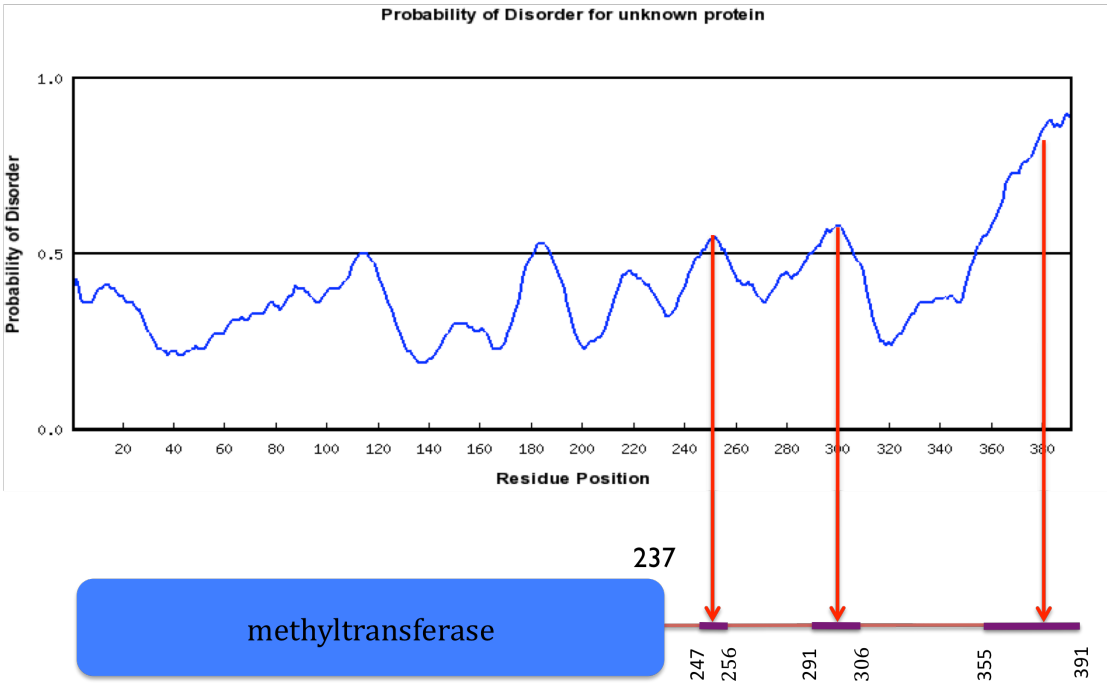



**Figure 13: Prediction of the potential disordered regions in dHen1 output by the RONN algorithm**

### 3.1.4.2. Secondary structure prediction and definition of constructs boundaries

Based on the data above and the secondary structure prediction programs Jpred and Phyre, with no predicted domain structure at the C-terminal region, constructs were designed for four C-terminal truncations of dHen1 by avoiding cutting into any predicted secondary structures such as α helix and β sheets. **Figure 14** shows the boundaries of the truncated constructs.

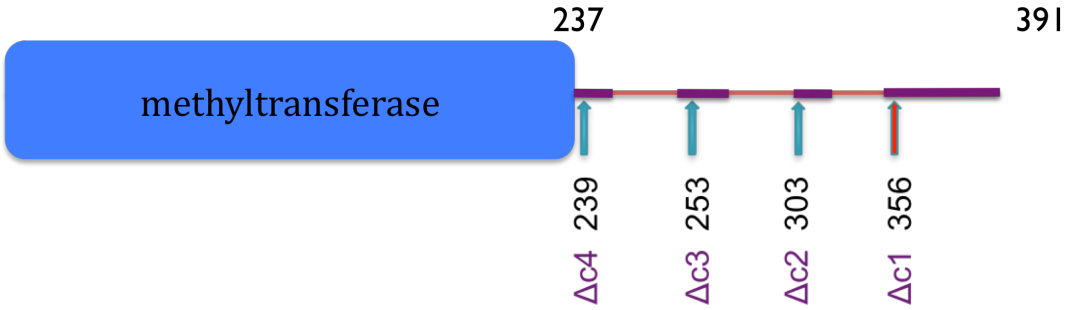



**Figure 14: Schematic diagram of C-terminal truncated dHen1 constructs boundaries**

### 3.1.4.3. Expression and solubility testing

The expression tests results for dHen1 constructs are summarised in **Table 7**. It showed that only the full length and ΔC1 (residue 1 – 356) constructs expressed solubly in *E. coli* system. The ΔC1 construct bound to the resin whereas, as before, the full-length protein failed to bind to the resin with N or C terminal tags. Thus we can hypothesize that the disordered region in the C-terminal tail (residues 357-371) inhibits interaction with the column, potentially through aggregation. In addition, it can be hypothesized (though it would be highly speculative) that the methyltransferase domain may not be an isolated domain and the C-terminal region of the protein (residues 238-356) may be making structural contacts with it and stabilizing it, see **Figure 15**.

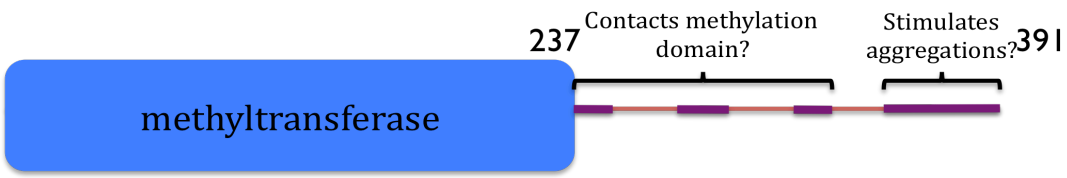



**Figure 15: C-terminal tail of dHen1.**

| Name of the constructs | Boundary | Tag | Solubility | Bind to Resin? |
|---|---|---|---|---|
| dHen1.1$_{(1-391)}$ | Full length | N-terminal His-tag | ✔ | ✘ |
| dHen1.2$_{(1-391)}$ | Full length | N-terminal GST-tag | ✔ | ✘ |
| dHen1.3$_{(1-391)}$ | Full length | C-terminal His-tag | ✔ | ✘ |
| dHen1.4$_{(1-356)}$ | ΔC1 | N-terminal GST-tag | ✔✔ | ✔ |
| dHen1.5$_{(1-356)}$ | ΔC1 | C-terminal His-tag | ✔ | ✔ |
| dHen1.6$_{(1-303)}$ | ΔC2 | N-terminal His-tag | ✘ | NA |
| dHen1.7$_{(1-303)}$ | ΔC2 | N-terminal GST-tag | ✘ | NA |
| dHen1.8$_{(1-303)}$ | ΔC2 | C-terminal His-tag | ✘ | NA |
| dHen1.9$_{(1-253)}$ | ΔC3 | N-terminal GST-tag | ✘ | NA |
| dHen1.10$_{(1-239)}$ | ΔC4 | N-terminal GST-tag | ✘ | NA |

**Table 7: Solubility testing results for dHen1 constructs**

### 3.1.5. Construct of dHen1.4$_{(1-356)}$ (ΔC1)

#### 3.1.5.1. Expression

Construct dHen1.4$_{(1-356)}$ (ΔC1) expression was further optimized by testing three different expression strains BL21(DE3) (Stratagene), Rosetta 2(DE3)pLysS (Merck Bioseciences), Arctic Express(DE3) (Stratagene)), two induction $OD_{600}$ (0.6, 0.9), two expression temperatures (12°C, 16°C) and two IPTG concentrations (0.1 mM, 0.5 mM). **Figure 16** shows the soluble fraction of each test expression condition and **Table 8** shows the expression condition corresponding to that in **Figure 16**. The bands (42 kDa in size) corresponding to dHen1.4$_{(1-356)}$ are indicated with an arrow.

| Lane | Expression Strain | Induction $OD_{600}$ | Induction Temperature | IPTG Concentration |
|---|---|---|---|---|
| 1 | Arctic Express (DE3) | 0.6 | 16°C | 0.1mM |
| 2 | Arctic Express (DE3) | 0.6 | 16°C | 0.5mM |
| 3 | Arctic Express (DE3) | 0.9 | 16°C | 0.1mM |
| 4 | Arctic Express (DE3) | 0.9 | 16°C | 0.5mM |
| 5 | Arctic Express (DE3) | 0.6 | 12°C | 0.1mM |
| 6 | Arctic Express (DE3) | 0.6 | 12°C | 0.5mM |
| 7 | Arctic Express (DE3) | 0.9 | 12°C | 0.1mM |
| 8 | Arctic Express (DE3) | 0.9 | 12°C | 0.5mM |
| 9 | Rosetta 2(DE3)pLysS | 0.6 | 16°C | 0.1mM |
| 10 | Rosetta 2(DE3)pLysS | 0.6 | 16°C | 0.5mM |
| 11 | Rosetta 2(DE3)pLysS | 0.9 | 16°C | 0.1mM |
| 12 | Rosetta 2(DE3)pLysS | 0.9 | 16°C | 0.5mM |
| 13 | Rosetta 2(DE3)pLysS | 1.2 | 16°C | 0.1mM |
| 14 | Rosetta 2(DE3)pLysS | 1.2 | 16°C | 0.5mM |
| 15 | BL21(DE3) | 0.6 | 16°C | 0.1mM |
| 16 | BL21(DE3) | 0.6 | 16°C | 0.5mM |
| 17 | BL21(DE3) | 0.9 | 16°C | 0.1mM |
| 18 | BL21(DE3) | 0.9 | 16°C | 0.5mM |
| 19 | BL21(DE3) | 1.2 | 16°C | 0.1mM |
| 20 | BL21(DE3) | 1.2 | 16°C | 0.5mM |

**Table 8: List of expression condition combinations for dHen1.4$_{(1\text{-}356)}$: expression strain, induction OD, expression temperature and IPTG concentration**

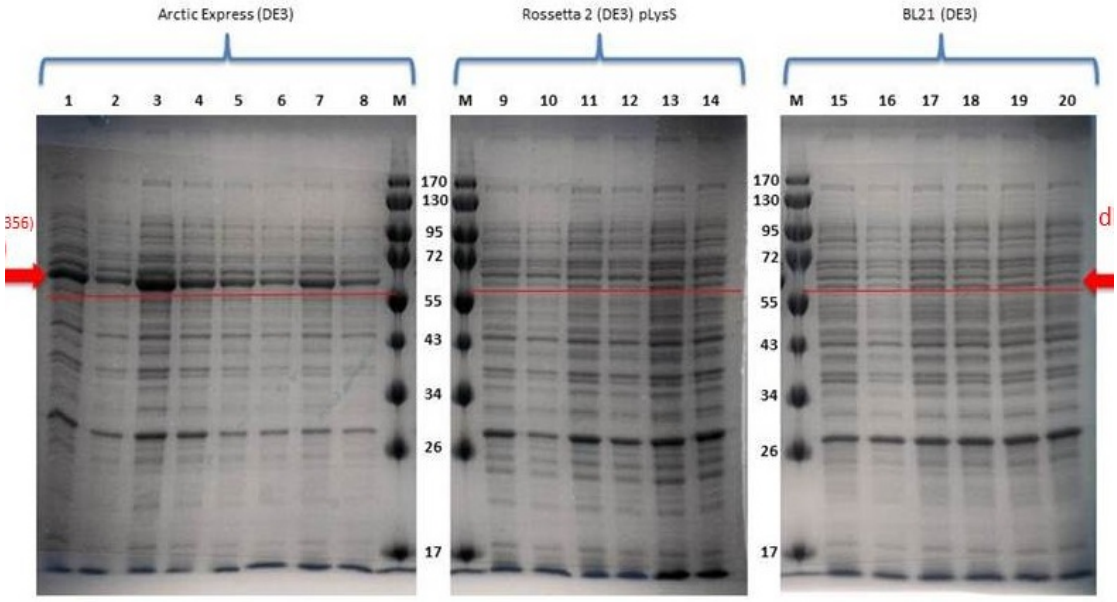



**Figure 16: SDS – PAGE gel of the soluble fractions for dHen1.4$_{(1-356)}$ expression test. The numbering across the top corresponds to the conditions listed in table 8. The red arrow indicates the band for dHen1.4$_{(1-356)}$**

**Figure 16** showed that dHen1.4$_{(1-356)}$ (42 kDa) protein expressed most solubly in the Arctic Express (DE3) strain and lane 3 showed the highest level of expression correlated to the optimal expression condition. Therefore the optimal expression condition was established as follows: Transformed *E. coli* cells (Arctic Express(DE3)) were grown at 37°C in LB till $OD_{600}$ = 0.9 and protein expression was induced by adding 0.1 mM IPTG at 16°C with overnight expression.

#### 3.1.5.2. Purification

The dHen1.4$_{(1-356)}$ (42kDa) protein purification scheme was developed de novo with final yield around 1 mg per 6L prep. It was purified in the sequence of GST-affinity column, 3C-cleavage to remove the GST-tag, mono-Q ion-exchange column and finally S75 size exclusion gel filtration column (Details in Chapter 2.) The purity of the protein after each step was examined by running the protein samples on 4-12% gradient SDS-PAGE gel and staining with coomassie blue dye.

After the initial GST-affinity column, the protein was contaminated with nucleic acid (indicated as high A260/280 value on Nanodrop spectrum) and still associated with chaperone proteins. The nucleic acid was separated after mono-Q column and a proportion of pure protein was separated from the pool that was still associated with chaperones. **Figure 17** summarizes the purification steps and it shows that most of the contaminant protein was separated from dHen1.4$_{(1-356)}$ after the mono-Q ion – exchange purification. The symmetrical elution peak of dHen1.4$_{(1-356)}$ on the gel filtration column and the single band of the correlated elution fractions on SDS-PAGE gel indicated that dHen1.4$_{(1-356)}$ with molecular weight of 42 kDa was a stable construct as shown in **Figure 18**.

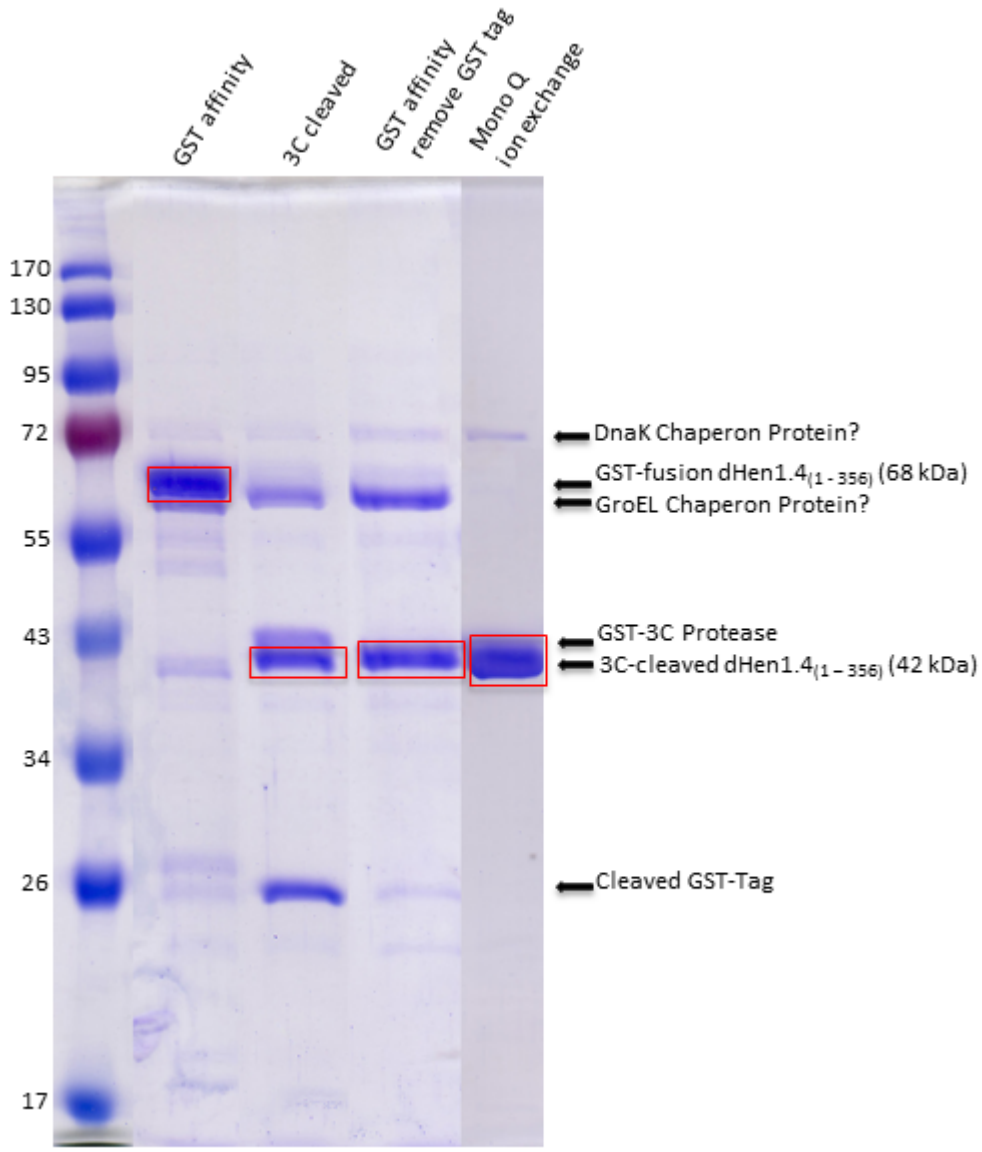



**Figure 17: SDS – PAGE gel result for step-wise purification scheme of dHen1.4$_{(1\text{-}356)}$. Each lane represents a purification step and the red box highlights the dHen1.4$_{(1\text{-}356)}$ proteins**

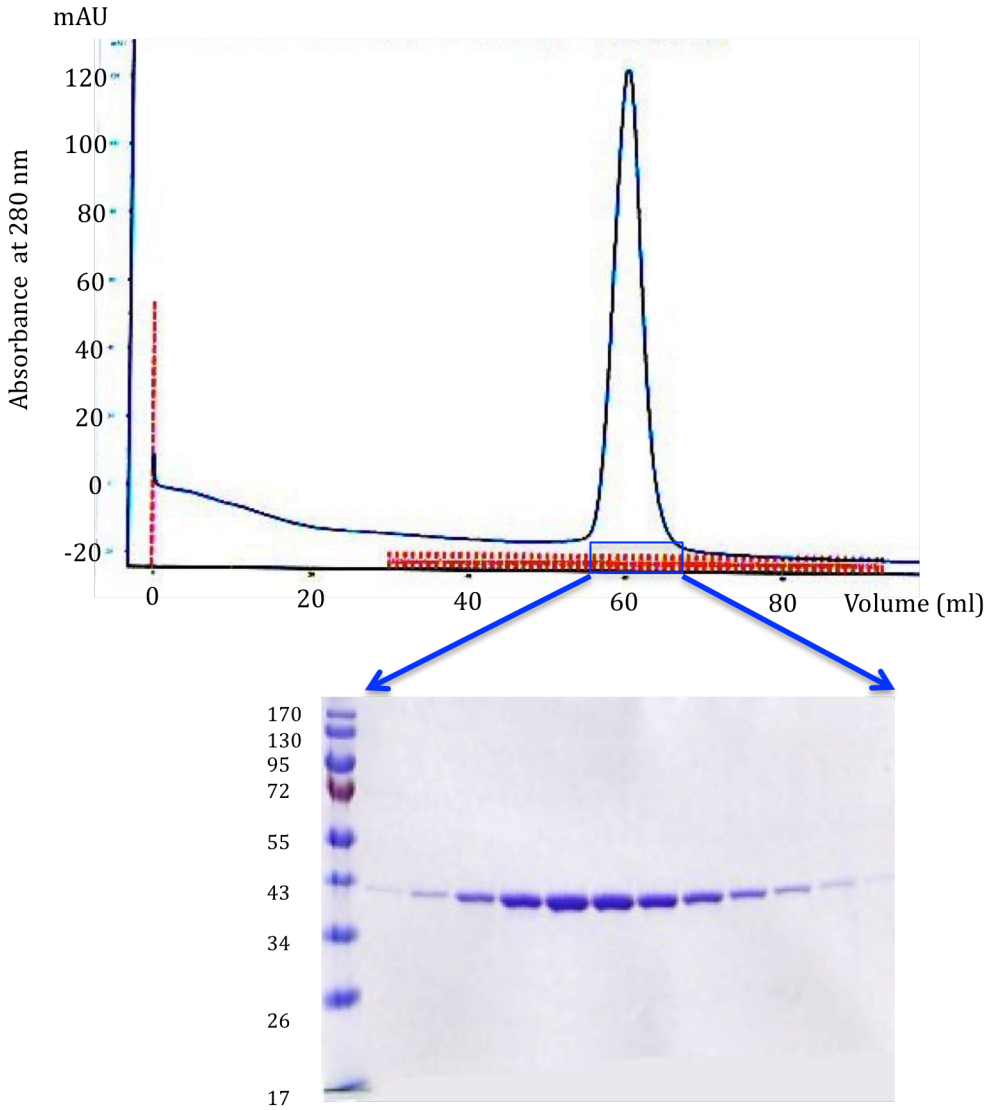



**Figure 18: dHen1.4$_{(1-356)}$ Gel filtration trace with SDS -PAGE gel of the eluted fractions.**

### 3.1.5.3. Limited Proteolysis

Limited Proteolysis is a technique for identification of exposed regions such as flexible loops on the protein backbone. The flexible regions of the protein have a tendency to hinder protein crystallization and a stable fragment produced from limited proteolysis is more likely to crystalize. Purified dHen1.4$_{(1-356)}$ was treated with trypsin and subtilisin separately with increasing concentration of the protease. The digestion reaction produced a smaller fragment with apparent

molecular weight of 28 kDa corresponding to that of the methyltransferase catalytic domain. **Figure 19** shows the result of the trypsin digest.

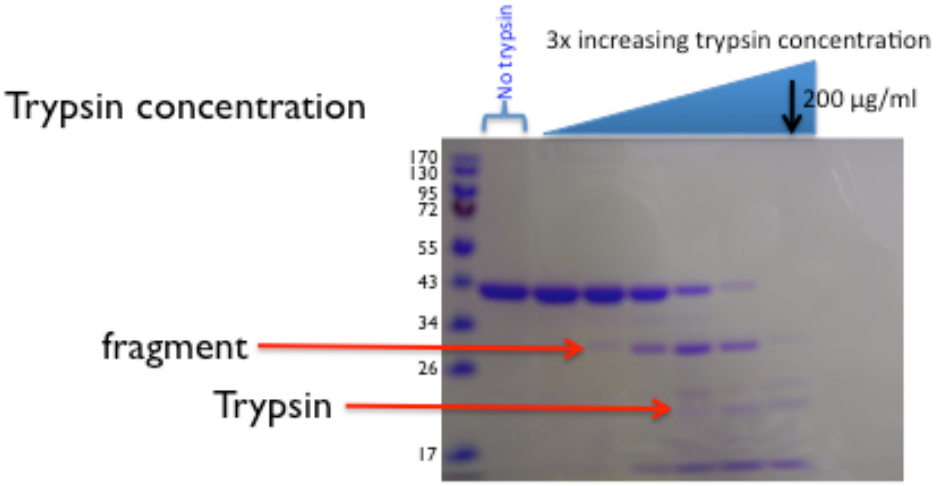



**Figure 19: result of a tryptic digest reaction of dHen1.4$_{(1\text{-}356)}$**

### 3.1.5.4. N-terminal sequencing

The N-terminal sequence of the fragment produced from proteolysis of dHen1.4$_{(1\text{-}356)}$ was SRASPL and the position of cleavage was subsequently confirmed. Unexpectedly the cleavage site was after the 98$^{th}$ residue arginine, which is in the middle of the catalytic domain shown in **Figure 20.**

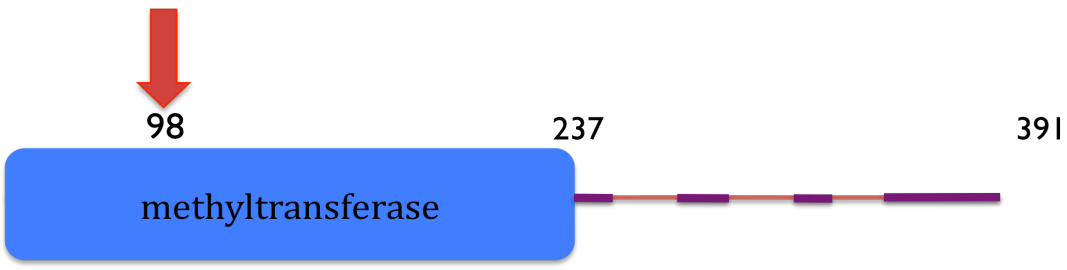



**Figure 20: Schematic representation of the tryptic sensitive loop on dHen1.4$_{(1\text{-}356)}$**

Therefore the limited proteolysis reaction showed that there is a tryptic sensitive loop in the catalytic domain of dHen1.4$_{(1-356)}$, but it did not produce a fragment that is suitable for designing a shorter construct.

#### 3.1.5.5. Crystallization trials

Extensive crystallization trials for dHen1.4$_{(1-356)}$ and complex of dHen1.4$_{(1-356)}$ with reaction product SAH (S-adenosyl-l-Homocysteine) were carried out at 4°C and 20°C in 96-well trays using Tecan and mosquito robots. The protein was concentrated to 200 μM (9.8 mg/ml) and the protein/SAH complex was formed in a molar ratio of 1:2. So far no protein crystals were obtained.

### 3.1.6. Revisiting full length dHen1

#### 3.1.6.1. Purification of full length dHen1 construct dHen1.2$_{(1-391)}$

dHen1.2$_{(1-391)}$ purification was carried out in the same way as described for dHen1.4$_{(1-356)}$. Though most of the protein in the cell lysate failed to bind to the GST resin, 300 μg of pure dHen1.2$_{(1-391)}$ protein was obtained from 6L LB culture.

#### 3.1.6.2. Limited Proteolysis

Limited proteolysis digestion for dHen1.2$_{(1-391)}$ with trypsin and subtilisin were carried out in the same manner as for dHen1.4$_{(1-356)}$ as shown in **Figure 21**. Full length protein could yield a longer proteolysis product.

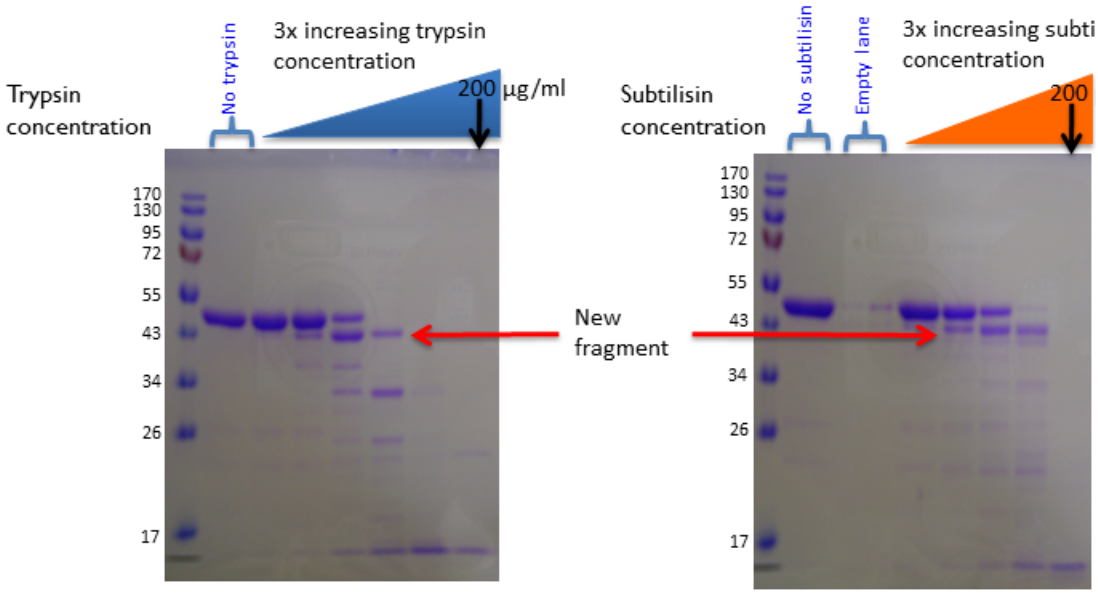



**Figure 21: The results of trypsin and subtilisin digest of full length dHen1.2$_{(1\text{-}391)}$**

The two proteolytic reactions both yielded an apparently stable fragment shown in **Figure 21**. The new fragment had a molecular weight that appeared to be slightly higher than that of dHen1.4$_{(1\text{-}356)}$; it indicated that potentially the fragment could be used to design a new stable construct.

### 3.1.6.3. N-terminal sequencing of limited proteolysis product

The new fragment shown in **Figure 22** from both types of limited proteolysis digestion of dHen1.2$_{(1\text{-}391)}$ full length protein was N-terminally sequenced.

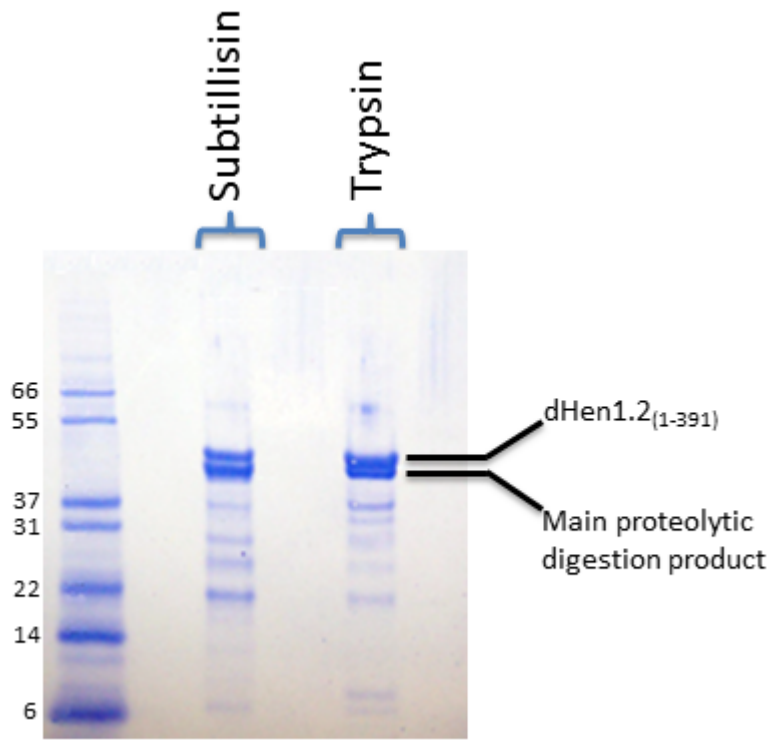



**Figure 22:Blot for N-terminal sequencing of trypsin and subtilisin digest product of dHen1.2$_{(1-391)}$**

The N-terminal sequence of the fragment from both trypsin and subtilisin digestion is GPLGSMF; which confirms that the fragment is a C-terminal truncation of the full length protein as the fragment shares the same N-terminal sequence as the full length dHen1.2$_{(1-391)}$.

### 3.1.6.4. Determination of the C-terminal truncation position

Mass spectrometry of limited proteolysis digestion of dHen1.2$_{(1-391)}$ products was carried out but was unable to reveal the molecular weight of the fragments. Therefore the position of the cleavage site needed to be confirmed by alternative methods.

As the fragment was the product of a tryptic digestion, the last residue of the fragment has either to be lysine or arginine. From SDS-PAGE gel the molecular weight of the fragment was in between dHen1.4$_{(1-356)}$ (41823 Da) and full length dHen1.2$_{(1-391)}$ (45981 Da) and near the 43 kDa marker, which indicated that the last residue of the new fragment lies between residues 357 and 391. There is only one lysine at K364, and two arginine residues at R368 and R384, that lie between residues 357 and 391 of dHen1 as shown in **Figure 23**.

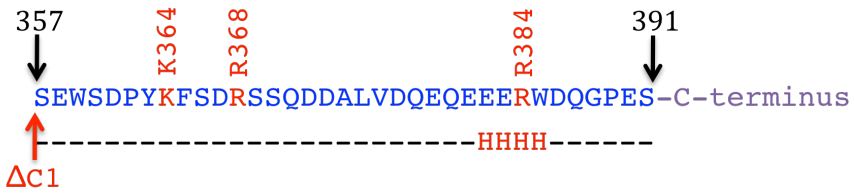



**Figure 23: the C-terminal tail sequence of dHen1 from residues 357 to 391 with secondary structure prediction by JPred**

JPred secondary structure prediction implied that the R384 was in a small helix turn. The calculated molecular weight for the fragment ending at K364 is 42816 Da, R368 is 43321 Da and R384 is 45181 Da. The fragment's band appeared more closely to the 43 kDa marker on a SDS-PAGE gel. Therefore the C-terminal truncation position appeared to be at either K364 or R368.

### 3.1.7. Testing Constructs ΔC5 - dHen1.12$_{(1-368)}$ and ΔC6 - dHen1.14$_{(1-364)}$

#### 3.1.7.1. Cloning, expression and purification of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$

Two new constructs were cloned, expressed and purified the same way as dHen1.4$_{(1-356)}$ based on the assumption that the C-terminal truncation position was at either position R368 or K364. **Figure 24** shows the schematic positions of

the new constructs and the trace of the S75 gel filtration column, with the eluted fractions on the SDS-PAGE gels for each construct respectively.

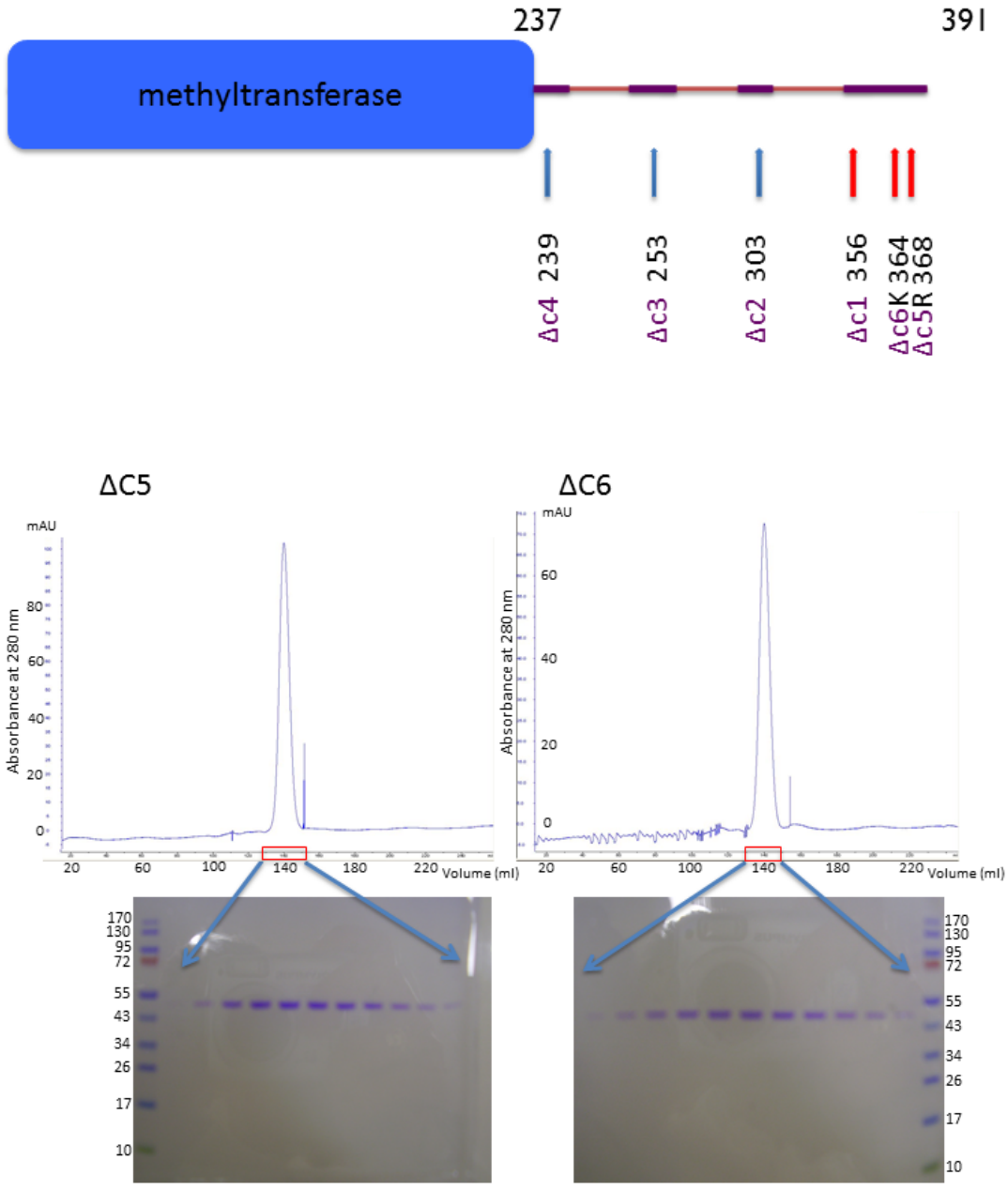



**Figure 24: Top image shows the summary of the constructs. Bottom images show the final purification step - gel filtration traces and the SDS-PAGE results of dHen1.12$_{(1\text{-}368)}$ and dHen1.14$_{(1\text{-}364)}$**

The trace and the single band on the SDS-PAGE gel showed that dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ were both pure and stable on gel filtration suggesting that both constructs were properly folded.

#### 3.1.7.2. Comparison with proteolysis

In order to determine which of the new constructs was the same as the digestion product, the purified sample of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ together with tryptic digest of full length dHen1.2$_{(1-391)}$ reaction products, dHen1.4$_{(1-356)}$ and full length dHen1.2$_{(1-391)}$ protein sample were run on a high resolution SDS-PAGE gel by only loading a minimal amount of 1 μl into a SDS-PAGE gel with wide lane. **Figure 25** shows the gel for comparing the sizes of digestion product with full length and C-terminal truncated construct of dHen1.

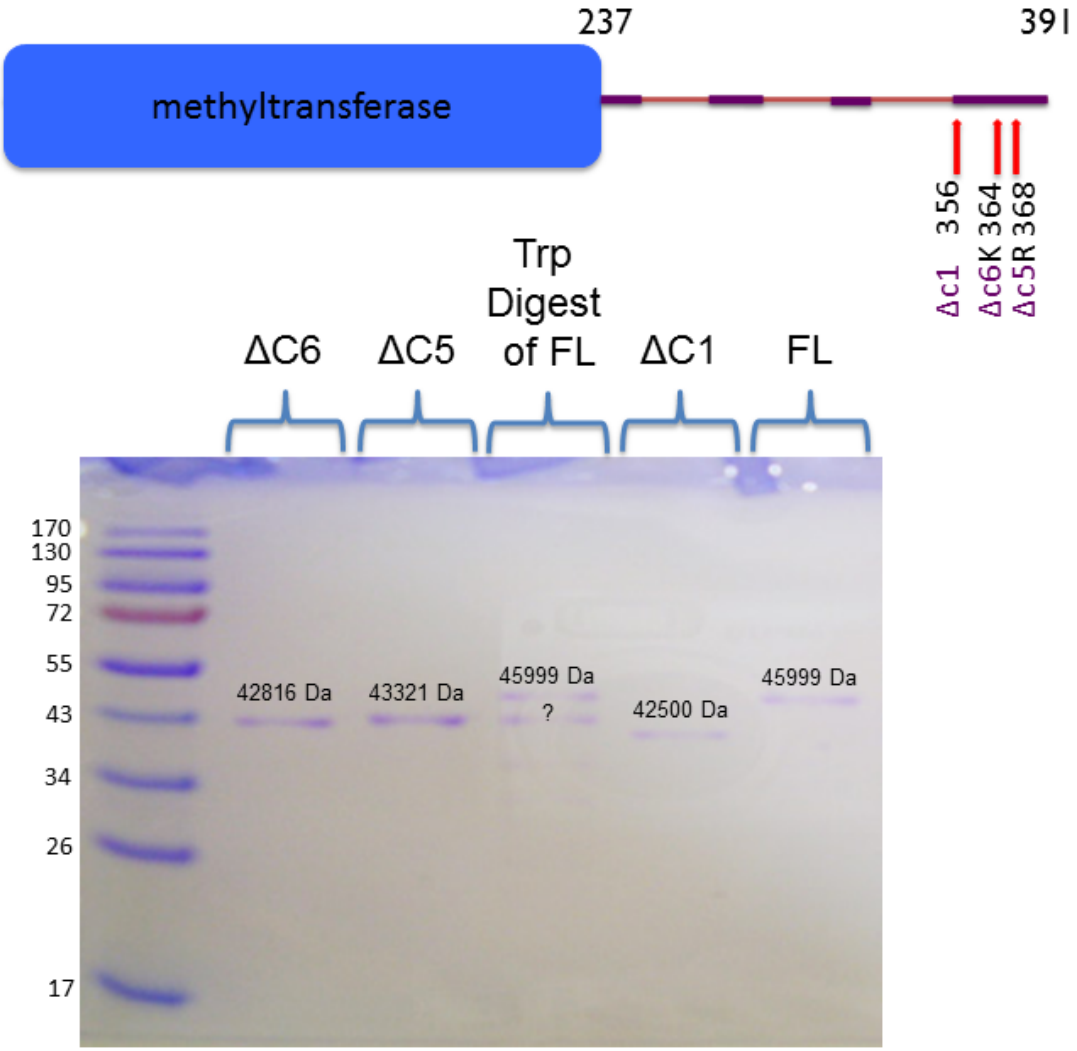



**Figure 25: Comparing the sizes of dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$**

As seen in **Figure 25**, the ΔC5 - dHen1.12$_{(1-368)}$ protein appeared to have the same molecular weight as the digested product of full length dHen1.2$_{(1-391)}$, which confirmed that the position for the C-terminal truncation based on proteolysis was R368.

### 3.1.8. Crystallization trials of the new constructs

The protein was concentrated to 200 μM (9.8 mg/ml). Protein/SAH/RNA mixture was formed in molar ratio of 1:2:3 and protein/SAH mixture was formed in molar ratio of 1:2. The following mixtures were put in to crystallization trials at 4°C in 24-well hanging drops screens:

- dHen1.12$_{(1-368)}$ / SAH/RNA -2mer
- dHen1.12$_{(1-368)}$ / SAH/ RNA -3mer
- dHen1.14$_{(1-364)}$ / SAH/ RNA -2mer
- dHen1.14$_{(1-364)}$ / SAH/ RNA -3mer
- dHen1.12$_{(1-368)}$ /SAH
- dHen1.14$_{(1-364)}$ /SAH

No protein crystals were obtained.

### 3.1.9. Discussion

In this section significant effort was made towards producing soluble and stable constructs of native dHen1 protein. Protein production has been optimized by choosing the most stable constructs under optimal expression conditions and purified dHen1 protein was believed to be suitable for structural and functional studies. However, as shown in section 3.1.5.2, during the entire protein dHen1.14$_{(1-364)}$ purification scheme, up to the final gel filtration step, there was chaperone protein presence, and this was also observed during the purification steps for both dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$. Hen1 activity studies are described in the next section.

## 3.2. Hen1 Mechanisms and Activity

### 3.2.1. Introduction

Various studies have been done on the animal Hen1 proteins to establish that Hen1 was the methyltransferase responsible for the 2'-O-methylation at the 3'-end of siRNAs in *Drosophila* and piRNAs across animal species (Horwich et al, 2007) (Saito et al, 2007) (Kirino & Mourelatos, 2007b) (Kamminga et al, 2010). It was also shown that in *vitro* animal Hen1 has substrate specificity for single stranded RNAs but not double stranded RNA duplexes. The structure of *Arabidopsis thaliana* HEN1, which by contrast binds double-stranded siRNA/miRNA duplexes, revealed that in addition to the methyltransferase domain at its C-terminus, there are four other domains involved in RNA binding at the N – terminus as shown in **Figure 26** (Huang et al, 2009).

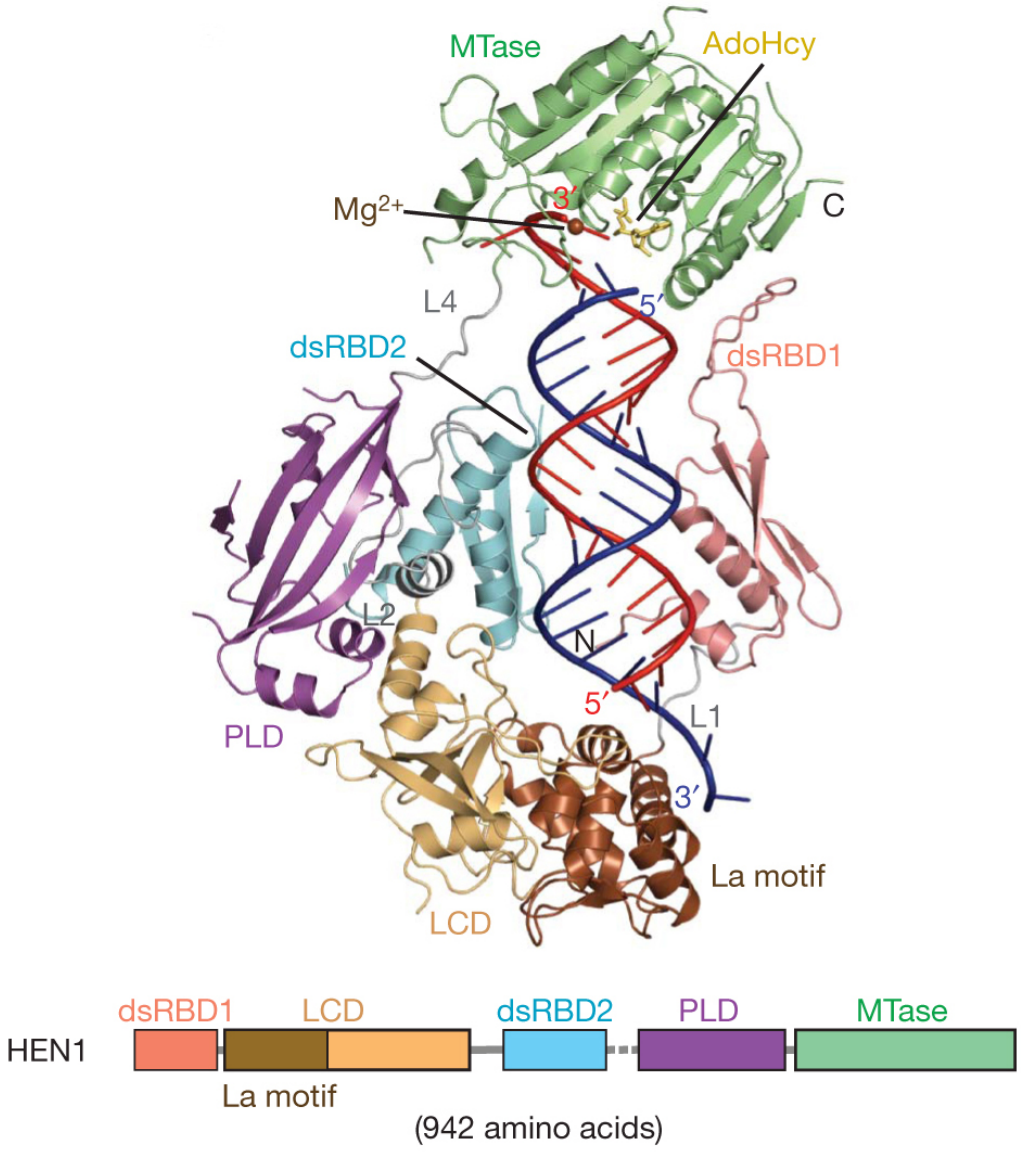



**Figure 26: Crystal structure of the Arabidopsis HEN1. (Huang et al, 2009) The domain structure is shown below with each individual domain coloured corresponding to the colouring of the domain in the crystal structure.**

The duplex is enveloped by dsRBD1 (double – stranded RNA Binding Domain 1), dsRBD2 (double-stranded RNA Binding Domain 2) and LCD (La- Motif Containing Domain) synergistically, and together these RNA binding domains provide specific substrate recognition and project the other terminus of the miRNA duplex towards the methyltransferase domain. The methyltrasferase

domain in turn can only efficiently coordinate the recognition and methylation of the 2nt – overhang within a limited range. Thus the HEN1 provide a mechanism for measuring the length of the substrate by limiting the distance between the N-terminal domains responsible for 5'- end capping and the C-terminal methyltransferase domain responsible for 3'- end methylation (Huang et al, 2009). Animal Hen1 homologues lack all the double stranded RNA binding domains that are present in *Arabidoposis thaliana* HEN1. Since there is no known motif or domain structure present in the C – terminal region of animal Hen1, it is therefore unknown if animal Hen1 could also measure the length of the RNA substrate as in *Arabidoposis thaliana* HEN1.

So far all in *vitro* methylation assays were performed on full length GST- fusion Hen1 proteins using single stranded RNAs ranging between 20nt to 40nt in length as substrates (Horwich et al, 2007) (Saito et al, 2007). Still, little is known about the mechanisms of Hen1 methylation and the reasons for the strict substrate specificity. RNA, especially single stranded RNA, is very unstable and for the purpose of crystallisation it is preferable to use shorter RNA to reduce the flexibility. Therefore, by testing the activity of dHen1 against different lengths of short RNAs, the stable minimal constructs for crystallisation trials could be defined.

In addition, there are no known domains or motifs present in the C – terminal regions of *Drosophila* Hen1 protein. However, as I showed in Section 3.1, the majority of the C –terminal region is required for the integrity of the Hen1 protein and for its soluble expression. Therefore it may interact directly with the methyltransferase domain.

In a pull down study, full length GST-fusion *Drosophila* Hen1 protein was shown to interact with the Piwi subfamily of Argonaute proteins from lysate (Saito et al, 2007). In zebrafish, it was shown that the GFP tagged full length zebrafish Hen1 co-localised also with Piwi subfamily of Argonaute proteins to Nuage (granular structures around the nucleus that are germ cell specific) (Kamminga et al, 2010). However no direct interaction was observed between the full length or C-terminal region of Zebrafish Hen1 with any part of the Piwi subfamily Argonaute proteins in the yeast two-hybrid experiment (Kamminga et al, 2010). Therefore having the stable constructs of *Drosophila* Hen1 protein could not only test its activity but also potentially explore the functionality of the C- terminal region of the enzyme.

The aim of the work in this section was to 1) determine the minimal functional constructs of dHen1 for crystallisation trials; 2) to examine any regulatory role of the C-terminal of the dHen1 protein to try to gain mechanistic insights into piRNA methylation and 3) if there are any length requirements for the RNA substrate by dHen1 in the same way as *Arabidoposis thaliana* HEN1 measures the length of miRNA duplex.

### 3.2.2. Methylation of short RNAs *in vitro*

#### 3.2.2.1. Analysis by autoradiography

The experiments were based on the in *vitro* methylation assay described in (Yu et al, 2005) but modified to suit the use of shorter RNAs. **Figure 27** shows the schematics of the experiment sequences.

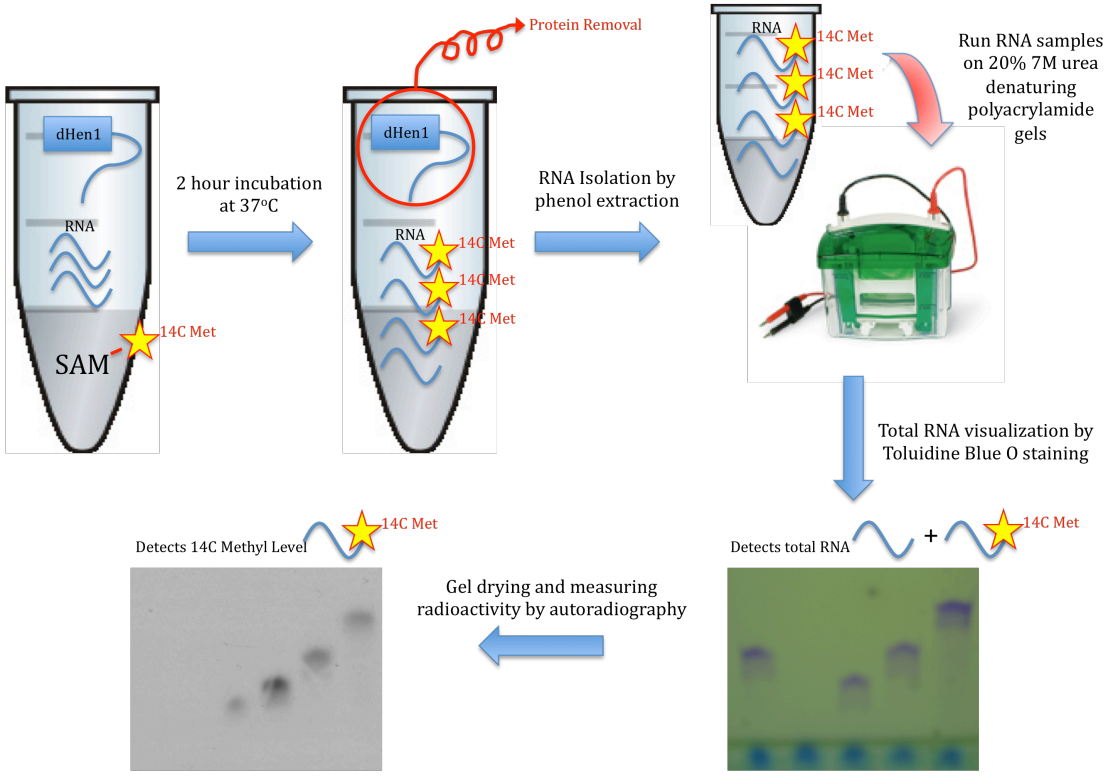



**Figure 27: Schematic representations of the in-vitro methylation assay and detection by autoradiography**

As dHen1 only methylates single stranded RNAs, the sequence of the RNA was carefully chosen using the DINAMelt melting profile prediction program to minimize the probability of RNA self-association. The 4 different lengths of RNA ranging from 5 nt to 22 nt listed in Table 4 (Chapter 2) were based on the sequence of let-7b micro-RNA.

The experiments were carried out at 26°C and 37°C. The melting temperature for 22 nt let-7b RNA duplex is around 27°C. Although 26°C is the physiological temperature for *Drosophila*, in order to prevent the self-annealing of the RNA, experiments were also carried out at 37°C to avoid artifacts as RNA duplex could not be methylated by dHen1 protein.

dHen1.2$_{(1-391)}$ designated as FL, dHen1.4$_{(1-356)}$ as ΔC1 and dHen1.12$_{(1-368)}$ as ΔC5 (domain structures shown in **Figure 28** below) were used in the methylation assay.

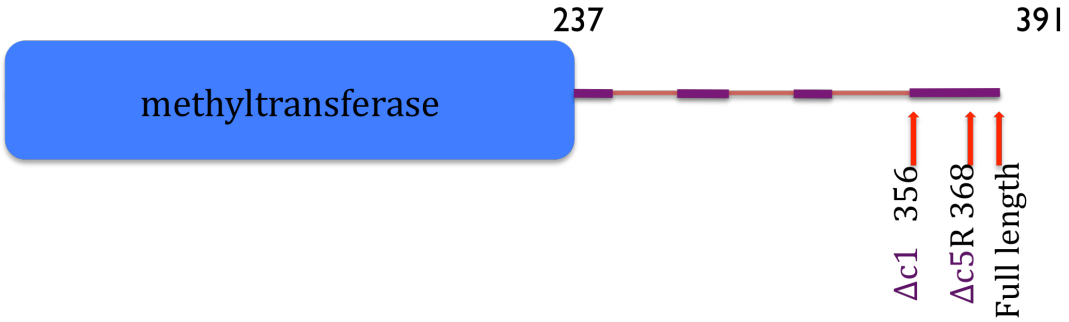



**Figure 28: Domain structure of dHen1.4$_{(1-356)}$ as ΔC1 and dHen1.12$_{(1-368)}$ as ΔC5**

Experiment 1 was carried out using FL and ΔC1 protein as a proof of principle and establishing optimal reaction conditions. **Figure 29** shows the results of the experiment. The Top table describes the conditions of each reaction mixture in the methylation assays including the incubation temperature, the length of the RNA substrates and the dHen1 constructs.

Experiment 1:

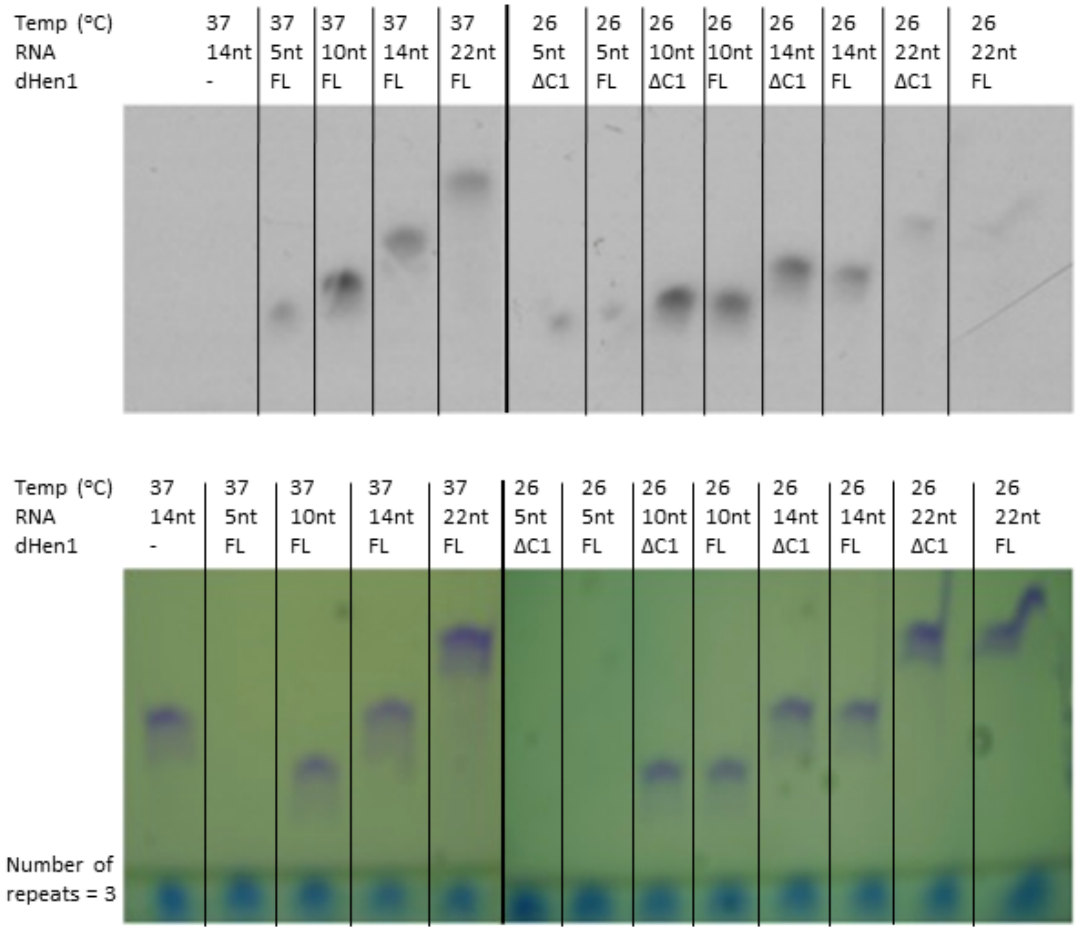



**Figure 29: dHen1 methylates single-stranded RNAs at 26°C and 37°C. The Top legend shows the reaction conditions including reaction temperature, length of the RNA substrate and the construct of dHen1 used in the experiment. The top result panel showed the film image by autoradiogrphy. The bottom result panel showed the TBO staining image of the acrylamide Urea gel.**

The top result panel of experiment 1 ($^{14}C$ autoradiograph of $^{14}C$-labelled RNAs) shows that 4 different lengths of single-stranded RNAs were methylated by dHen1 protein and its truncated constructs. The intensity of the band represented the level of RNA methylation, which is proportional to the activity of the dHen1 protein. The bottom panel shows the staining images (TBO, Toluidine Blue O) of the top gels, which indicates that the same amount of single-stranded RNA was used in all the lanes except for 5 nt RNA, which did not show up in TBO staining.

In Experiment 1, the negative control (the first lane from the left) had no enzyme present confirming that there was no non-specific incorporation of $^{14}C$ to RNA. At 26°C the low level of activity of dHen1 against 22 nt RNA could be explained by self-annealing of the RNA into double-stranded RNA, hence preventing the enzyme from methylating it. Such an artifact was overcome by carrying out the experiment at 37°C and the activity of the enzyme was not compromised at 37°C either as shown in the first 5 lanes of the **Figure 29**.

The result showed that both dHen1 constructs exhibit higher activity against 10 nt and 14 nt RNA substrate comparing to the 22 nt natural length substrate. The 10 nt RNA appeared to be a slightly better substrate than 14 nt RNA. In addition the shorter C-terminal truncated construct ΔC1 dHen1.4$_{(1\text{-}356)}$ showed consistently higher activity against all lengths of RNA substrate compared to that of the full length dHen1.2$_{(1\text{-}391)}$.

Experiment 2 subsequently incorporated ΔC5 to further evaluate the effect of various truncated dHen1 constructs. As in experiment 1 the top table shows the conditions for each of the reaction mixtures and **Figure 30** shows the results of experiment 2.

Experiment 2:

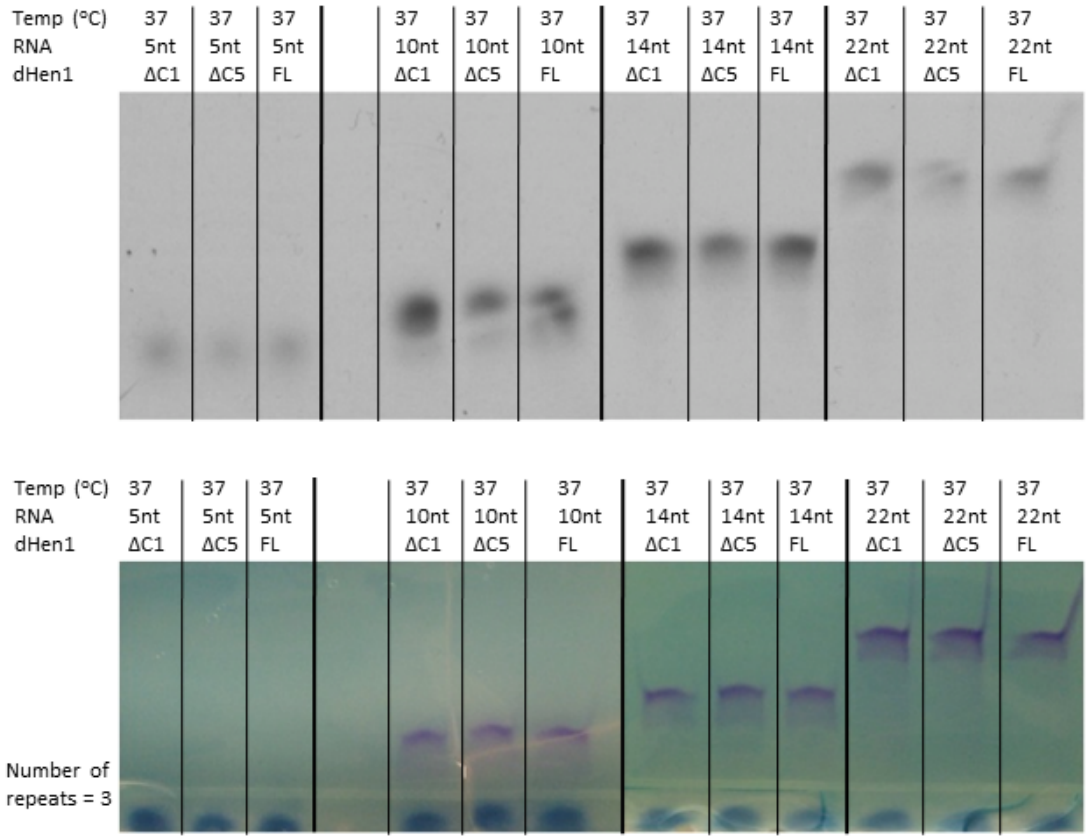



**Figure 30: dHen1 FL and truncated constructs methylate single -stranded RNAs at 37°C. The top legend shows the reaction conditions including reaction temperature, length of the RNA substrate and the construct of dHen1 used in the experiment. The top result panel showed the film image by autoradiogrphy. The bottom result panel showed the TBO staining image of the acrylamide Urea gel.**

The results of experiment 2 together with results of experiment 1 demonstrated that dHen1 and its truncated constructs had the highest level of activity against 10 nt and 14 nt RNA.

To summarize, shorter single stranded RNAs that are shorter than the full-length piRNAs serve as better substrates for methylation by Hen1 proteins and truncated constructs of Hen1 have higher enzymatic activities.

The caveat of the result is that the 5 nt RNA did not show up in the TBO gel staining stage possibly due to inefficient covalent fixing of the small RNA by acetic acid to the polyacrylamide gel. Thus the apparent low incorporation of the $^{14}C$ – methyl group to the 5 nt RNA could be a direct result of it. Alternatively, the 5 nt RNA could stain inefficiently. This effect can be further examined using a different detection method such as liquid scintillation counting.

#### 3.2.2.2. Analysis by scintillation counting

To further confirm the results from the methylation assay and to quantitatively compare the activities of various Hen1 constructs versus different RNA substrates, liquid scintillation counting was performed on both ΔC1 dHen1.4$_{(1-356)}$ and ΔC5 dHen1.12$_{(1-368)}$ using the same RNA oligos as substrates as in previous methylation assays. **Figure 31** shows the schematic sequence of the experiment.

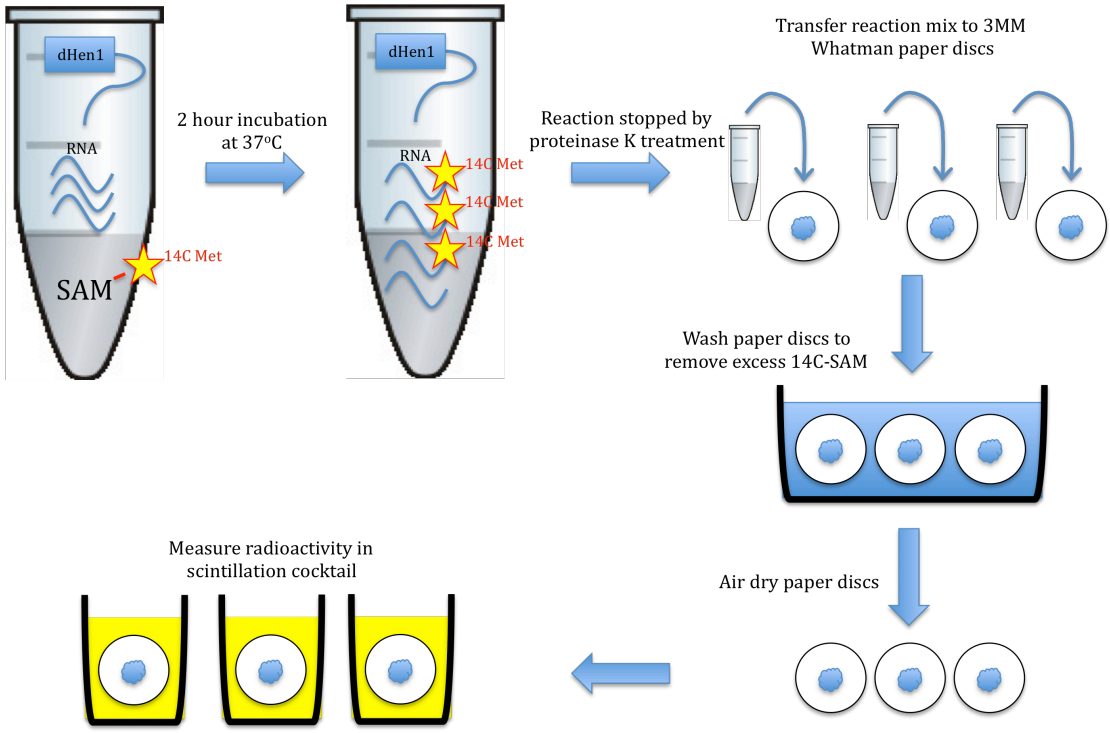



**Figure 31: schematic representations of the in *vitro* methylation assay and detection by liquid scintillation counting.**

Three control experiments were also carried out at the same time of the reaction to monitor the non specific signal from S-adenosyl-L-(methyl-$^{14}$C) methionine. In the 10mer RNA control dHen1 protein was omitted from the reaction mix; in the dHen1 control experiment RNA was omitted and in the SAM control experiment both dHen1 protein and RNA substrate were omitted. **Figure 32** shows the results of the scintillation counting.

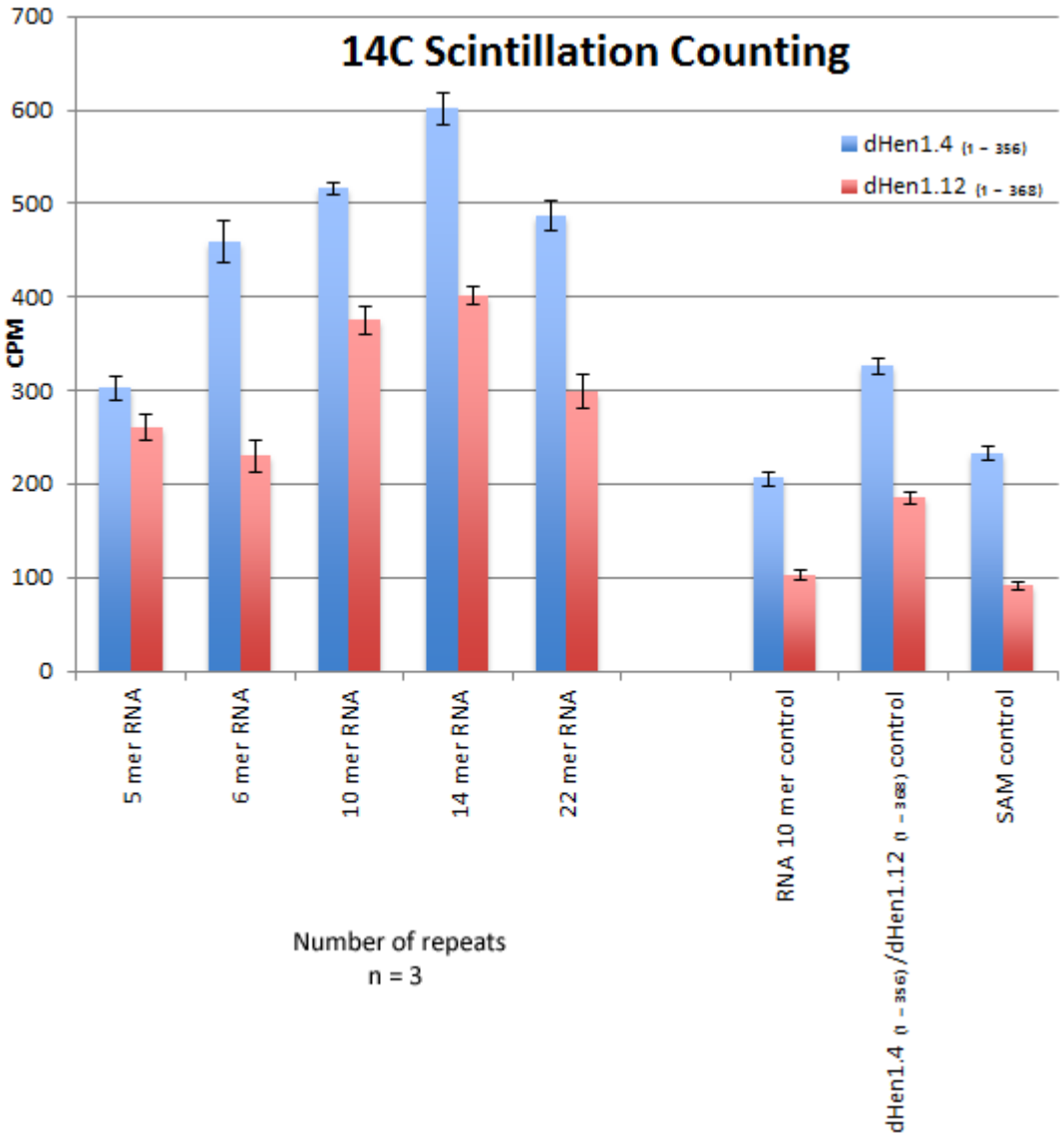



**Figure 32: result of the liquid scintillation counting using two dHen1 constructs against 5 different lengths of small RNA substrates. The control experiment results are shown on the three right hand side column.**

As indicated in **Figure 32**, consistent with previous findings, both dHen1.4$_{(1\text{-}356)}$ and dHen1.12$_{(1\text{-}368)}$ showed the highest activity against 10 and 14 mer RNA. Both constructs showed a similar level of activity against short 5/6 mer RNA and long 22 mer RNA, which was also demonstrated in the previous methylation assays.

### 3.2.3. Interaction with Argonaute

Both Hen1 proteins and Piwi subfamily of Argonaute proteins interact with piRNA, more specifically both Hen1 proteins and PAZ domain of the Piwi Argonaute proteins interact with the 3'-end of the piRNA. In principle, Hen1 proteins could dislodge the unmethylated 3'-end of the piRNA from the PAZ domain in order to methylate it and allow maturation of the piRNA. Therefore there could be an interaction between the Hen1 protein and the PAZ domain of Piwi subfamily of Argonaute protein.

To examine if there is a direction interaction between full length dHen1 and the PAZ domain of a Piwi subfamily of Argonaute protein, three runs of analytical gel filtration experiments were performed using full length *Drosophila* Hen1 and PAZ domain from Drosophila Aubergine, a *bona fide* Hen1 substrate. Firstly 15 μM of dHen1.2$_{(1\text{-}391)}$, secondly 15 μM of AubPAZ3$_{(279\text{-}398)}$ (Section3.3 and 3.4) and finally the two proteins were mixed in a 1:1 molar ratio at a final concentration of 15 μM and run on the analytical Superdex75 gel filtration column. **Figure 33** shows the trace of the runs

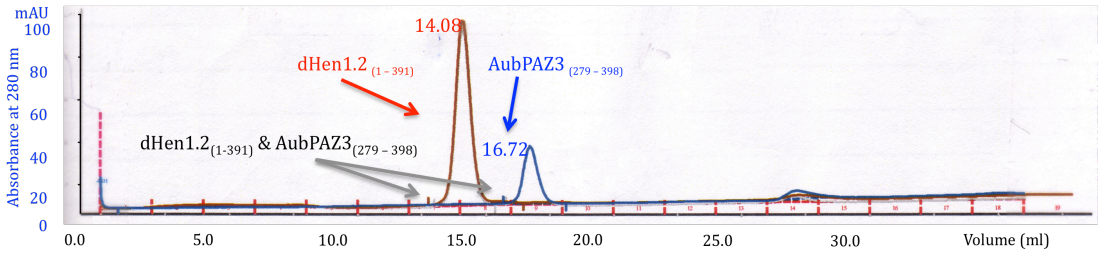



**Figure 33: Analytical Gel filtration run for dHen1.2$_{(1\text{-}391)}$ and AubPAZ3$_{(279\text{-}398)}$. The run for the protein mixture is shown in grey colour, which completed superimposed with the individual runs.**

As shown in **Figure 33** the final run peak superimposed with the individual run and there were no indications of complex formation on the gel filtration column

at 15 μM concentration. This suggest that there is no strong interaction between the PAZ domain of the Piwi subfamily protein Aubergine and full length *Drosophila* Hen1 protein, at least not detectable on a gel filtration column.

### 3.2.4. Impact of C- terminal tail on confirmation

The methylation assays earlier showed that there were potential mechanistic effects of the C – terminal tail of the dHen1 protein. To investigate this further, dHen1.2$_{(1-391)}$ (FL), dHen1.4$_{(1-356)}$ (ΔC1), dHen1.12$_{(1-368)}$ (ΔC5) and dHen1.14$_{(1-364)}$ (ΔC6) were run on the S200 analytical gel filtration column and **Figure 34** shows the trace.

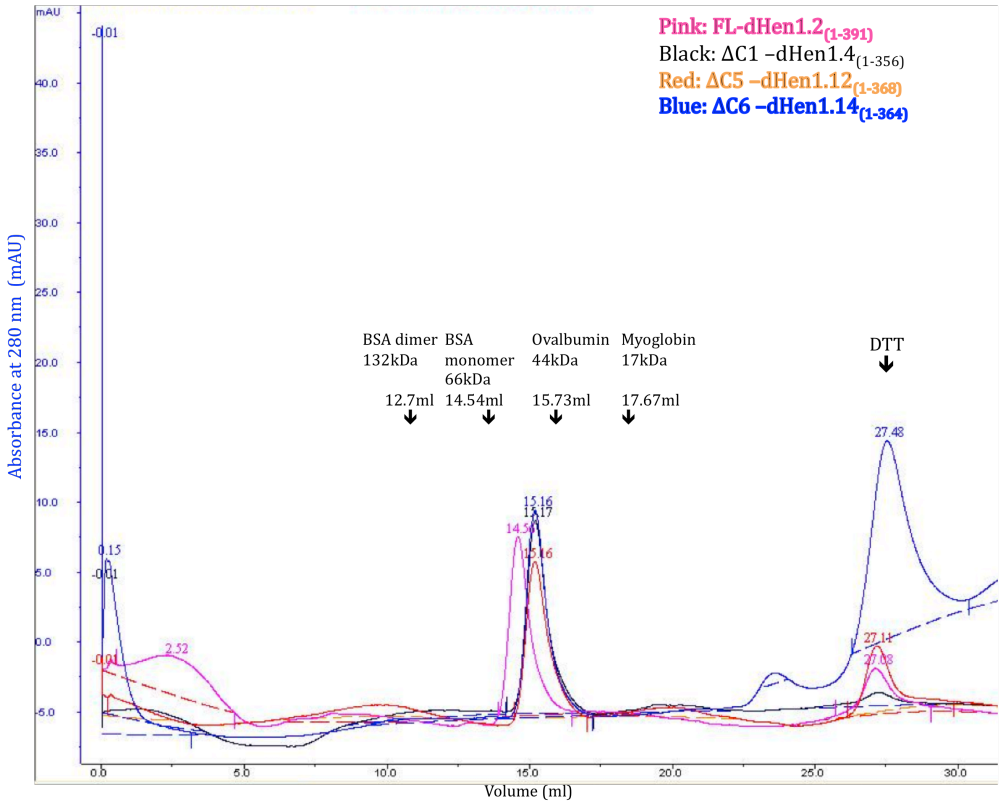



**Figure 34: trace of analytical gel filtration for dHen1.2$_{(1-391)}$, dHen1.4$_{(1-356)}$, dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$. Positions of the markers with respective molecular weight and elution volume are also shown.**

The analytical superdex200 column calibration is shown in **Table 9** and **Figure 35**.

| Protein Marker | MW (Da) | $Log_{10}$ MW | Ve (ml) | (Ve - Vo)/(Vt - Vo) | |
|---|---|---|---|---|---|
| BSA Dimer | 132000 | 5.12 | 12.7 | 0.312 | Vo = 7.57 ml |
| BSA monomer | 66000 | 4.82 | 14.54 | 0.424 | Vo = Void Volume |
| Ovalbumin | 44000 | 4.64 | 15.73 | 0.497 | Vt = 24 ml |
| Equine Myoglobin | 17000 | 4.23 | 17.67 | 0.615 | Vt = Column Volume |
| Vitamin B12 | 1350 | 3.13 | 20.87 | 0.809 | |

**Table 9: List of protein standards for gel filtration column calibration**

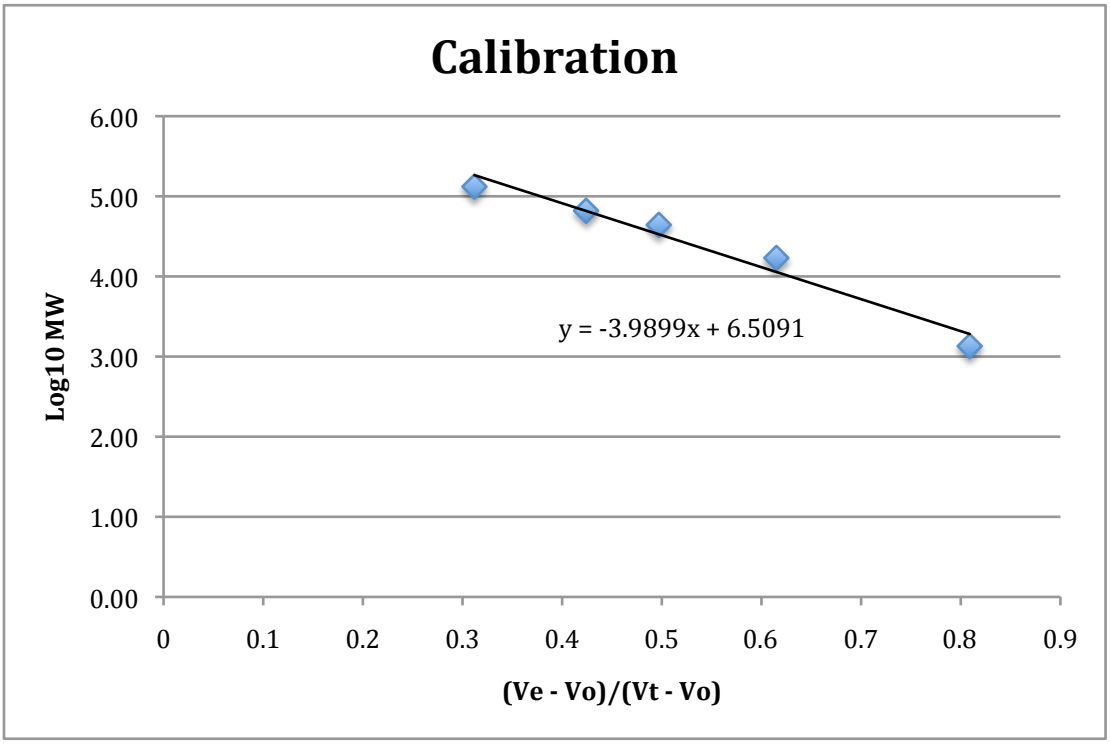



**Figure 35: calculation of molecular weight from the elution volume based on the calibration.**

From the calibration the molecular weight of the protein can be calculated from the elution volume.

Line of best fit: y = -3.9899x + 6.5091

For calculated molecular weight:

$$\text{Log}_{10}\ MW = -3.9899\ x + 6.5091$$

$$MW = 10^{(-3.9899\ x + 6.5091)}$$

where $x = (Ve - Vo) / (Vt - Vo)$

**Table 10** lists the elution volume of four different dHen1 constructs and their respective calculated molecular weight.

| Construct | Molecular Weight | Elution Volume Ve | (Ve – Vo) / (Vt – Vo) | Calculated MW |
|---|---|---|---|---|
| FL - dHen1.2$_{(1 - 391)}$ | 45.98 kDa | 14.56 ml | 0.4254 | 64.84 kDa |
| ΔC5 - dHen1.12$_{(1 – 368)}$ | 43.43 kDa | 15.16 ml | 0.4620 | 46.36 kDa |
| ΔC6 - dHen1.14$_{(1 - 364)}$ | 42.82 kDa | 15.16 ml | 0.4620 | 46.36 kDa |
| ΔC1- dHen1.4$_{(1 - 356)}$ | 41.81 kDa | 15.17 ml | 0.4625 | 46.08 kDa |

**Table 10: analytical gel filtration results**

The analytical gel filtration results showed that the full length dHen1 protein runs anomalously on the column, whereas the C-terminal truncated constructs run very similar to their expected molecular weight. This result could mean that the full length dHen1 associates weakly via its C-terminal residues, causing it to migrate more slowly. The effects on dispersity could explain the difficulty to purify the full length protein on a Ni- NTA column to trap the 6-Histidine tag.

### 3.2.5. Discussion

In order to study dHen1 methyltransferase structurally, various constructs of dHen1 including full length and C-terminal truncated versions were cloned, expressed and purified. There are three constructs (dHen1.4, dHen1.12 and dHen1.14) that appeared to be stable and the purity was confirmed by running

the SDS-PAGE gel. Several unsuccessful attempts have been made towards crystallizing the apo-protein, protein/SAH and protein/SAH/2 nt or 3 nt RNA complex.

The methylation assay has revealed that the protein was most active with 10nt or 14 nt RNA as the substrates as opposed to with 5 nt RNAs. This result could mean that 10 nt or 14 nt RNA were the optimal length for forming a complex with dHen1 protein and dHen1 interact with RNAs that are greater than 5 nt in length. The initial assay result also indicated that the activity of the C-terminal truncated proteins was not abolished or even hampered by the truncation, which confirmed that the truncated constructs are suitable for crystallization trials.

In addition, the consistently higher level of activity exhibited by the truncated dHen1.4$_{(1\text{-}356)}$ indicated that the C-terminus of dHen1 protein may play a role in the regulation of the methylation reactions, potentially through inhibition of the catalytic domain. The confirmation would be altered upon interaction with the RISC complex. Though, when a gel filtration binding experiment was performed with full length *Drosophila* Hen1 and the PAZ domain from a Piwi subfamily protein, no interaction was observed.

**Figure 36** summaries the results of Section 3.1 and Section 3.2 for the mechanism for Hen1 substrate recognition.

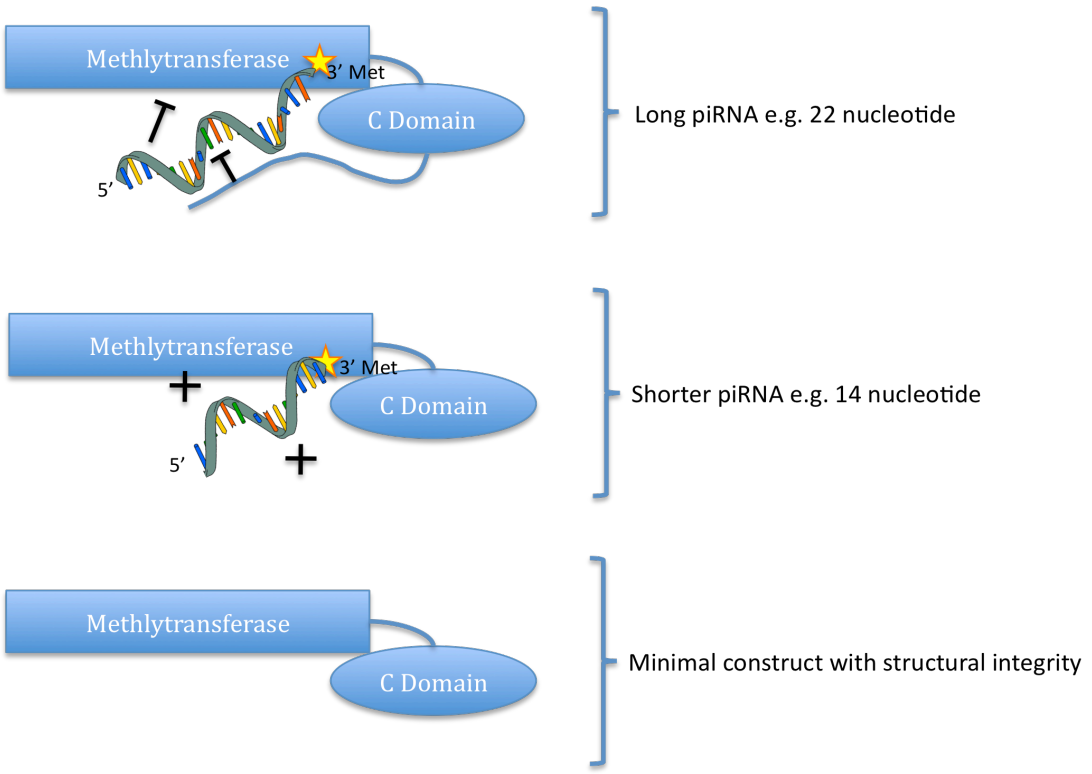



**Figure 36: schematic summary of the working hypothesis for the Hen1 protein mechanism. The top panel shows that longer RNA / full length C –terminal tail of Hen1 protein inhibit the methylation activity. The middle panel shows that a shorter Hen1 construct exhibit higher activity against shorter RNA substrates. The bottom panel shows there is a minimal length of C –terminal tail required for the Hen1 protein structural integrity.**

Together with the solubility results of various dHen1 C-terminal truncation constructs and the methylation activity assay, it is plausible to hypothesize that the N-terminal methyltransferase domain is not an isolated domain, but making structural contact with C-terminal part of the protein and retaining the enzyme in a relatively inactive state. Upon removing the extreme end of the C-terminal tail, the catalytic domain is more exposed and the enzyme becomes more active. In order for dHen1 to be fully activated, additional interaction partners may be required and protein could undergo more significant conformational changes. Inhibition by the C-terminal tail could prevent unwanted methylation reactions in the cytoplasm until Hen1 encounters the authentic substrate that is bound to the Piwi subfamily of Argonaute protein in RISC. Unlike *Arabidopsis thaliana*

HEN1, there is no 22 nt length mearsurement. This conclusion (though speculative) summarised in Figure 36 could be used for a future working hypothesis to be tested.

## 3.3.PAZ – methyl piRNA structure: cloning, purification, crystallisation and structure determination

### 3.3.1. Introduction

piRNAs, methylated at the 3'- ends, associate only with Argonaute proteins from the Piwi subfamily. In Human, there are four Argonaute proteins belong to the Piwi subfamily – Hili, Hiwi1, Hiwi2 and Hiwi3; in Mouse there are three – Mili, Miwi and Miwi2; in Zebrafish there are also two – Ziwi and Zili; in *Drosophila,* where most of the studies on piRNAs are done, there are three Piwi Argonautes – Aubergine (Aub), Piwi and Ago3 (Cox et al, 1998). Two of the Argonaute proteins' domains are primarily responsible for small RNAs binding: the MID domain has a 5' – phosphate group binding site (Parker et al, 2004) and the PAZ domain is responsible for the 3' end binding (Lingel et al, 2003) (Song et al, 2003).

Structurally, there is no information on full length eukaryotic Argonaute proteins, as they are difficult to express solubly. The majority of the structural studies have been conducted on the prokaryotic Argonaute proteins (Song et al, 2004) (Parker et al, 2005) (Wang et al, 2009b). The only eukaryotic Argonaute protein structures available are the PAZ domain of the Ago subfamily Argonaute proteins (*Drosophila* Ago1, *Drosophila* Ago2, human Ago1) (Ma et al, 2004) (Song

et al, 2003) (Yan et al, 2003) (Lingel et al, 2003) and the MID domain of the Ago subfamily Argonaute protein QDE-2 from Neurospora crassa (Boland et al). So far there is no structural information on any of the Piwi subfamily of Argonaute proteins. **Figure 37** summarises the currently known structures of Argonautes.

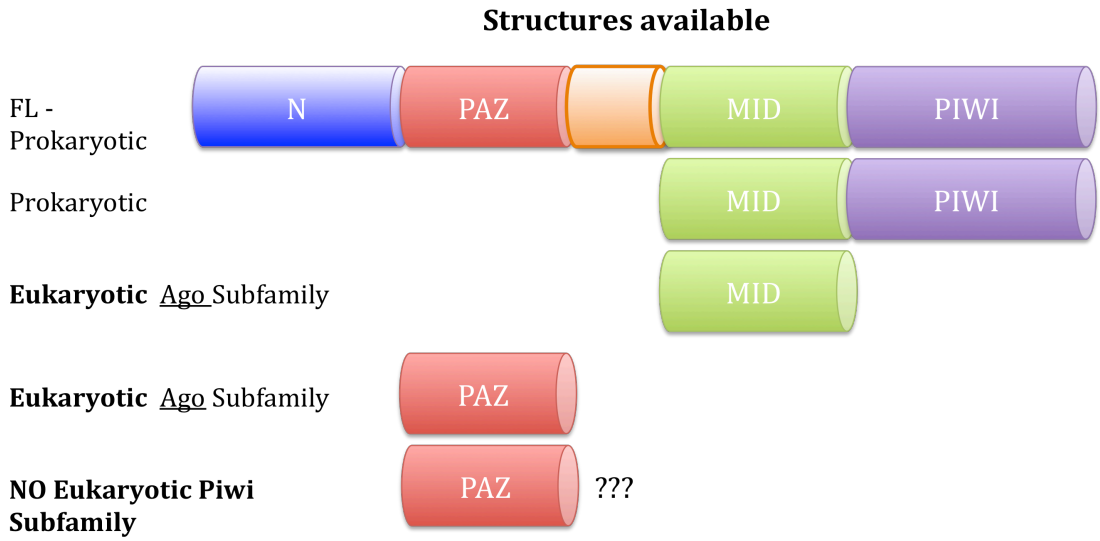



**Figure 37: Schematic summary of currently available Argonaute protein structures**

Key residues that form the 3'-end binding pocket and are involved in the 3'- end recognition, determined from the structure studies of the Ago subfamily of PAZ domain, are conserved across the Ago and Piwi subfamilies. However, the two subfamilies of Argonaute proteins have a low level of overall sequence homology and especially the PAZ domains of Piwi and Ago subfamilies have very diverse sequences as shown in alignment **Figure 38**.

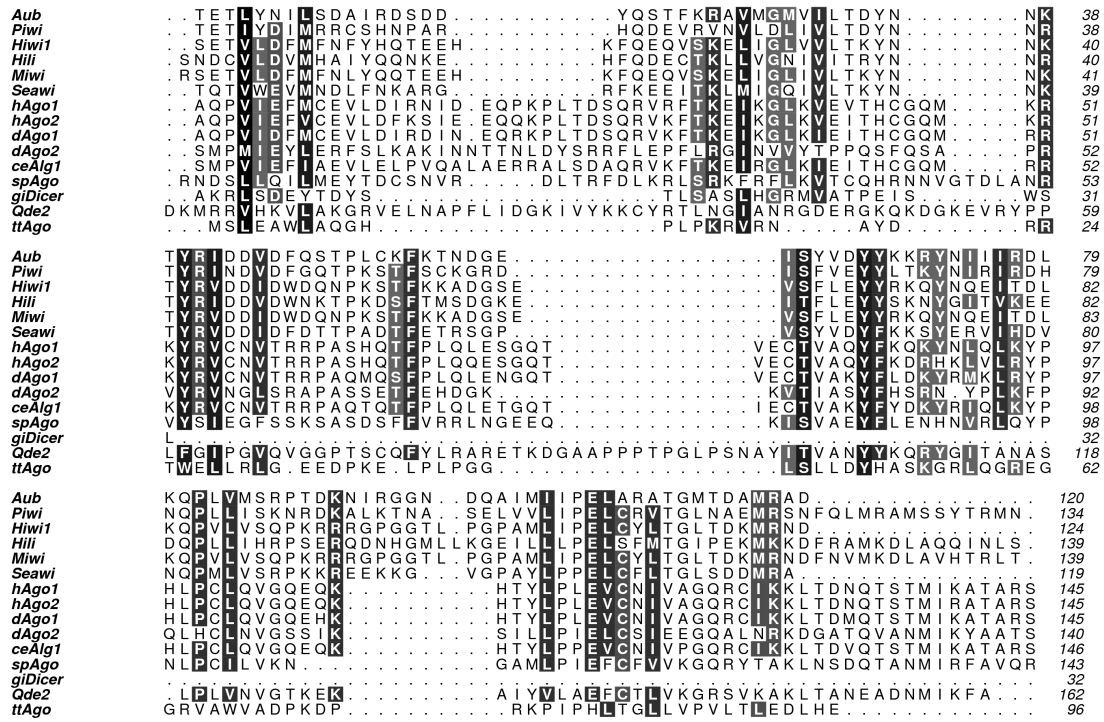



**Figure 38: Sequence alignment of Argonaute protein PAZ domains across different subfamilies including Aubergine, Piwi, Hiwi1, Miwi and Seawi from the Piwi subfamily, hAgo1, hAgo2, dAgo1 and dAgo2 from the Ago subfamily.**

Previous studies showed that methylation inhibits binding to the Ago subfamily PAZ domain (Ma et al, 2004), and yet the Piwi subfamily of Argonaute proteins can correctly recognise and load piRNAs with a 2'-O – methyl modification at the 3' end. This raises the question of how do Piwi Argonaute proteins, in particular the PAZ domain of the Piwi Argonaute proteins, recognise and accommodate the methyl group at the 3' end of piRNAs? Do they bind methylated 3'- ends? The aims of the studies described in the current and the next section (3.3 and 3.4) were to provide: firstly, the first structural information on any part of Piwi subfamily of Argonaute proteins; secondly, the first PAZ domain structure of Piwi subfamily of Argonaute proteins which in terms of sequence identity, are divergent from the Ago-subfamily of PAZ domains. In addition, through the production and structure determination of a PAZ domain of Piwi subfamily of Argonaute protein in complex with its natural substrate, to address the following

questions: 1) How do the Piwi Argonaute protein, in particular the PAZ domain recognise the 2'-O-methyl modification on piRNAs on a molecular level; 2) what is the difference between Piwi and Ago subfamily of Argonaute protein PAZ domain in recognition of methylated 3'-end? Aubergine PAZ domain from *Drosophila* was chosen as the candidate for structure studies as most existing biochemical studies have been done on *Drosophila*, and clones are available from *Drosophila* Genomic Resources Centre.

### 3.3.2. Construct design

The PAZ domain of Piwi and Ago subfamily of Argonaute proteins are very diverse in sequence. Based on the secondary structure prediction and the sequence alignment, four constructs (shown in **Figure 39**) for the Aubergine PAZ domain proteins were designed by avoiding cutting into any predicted secondary structures such as α helix and β strands.

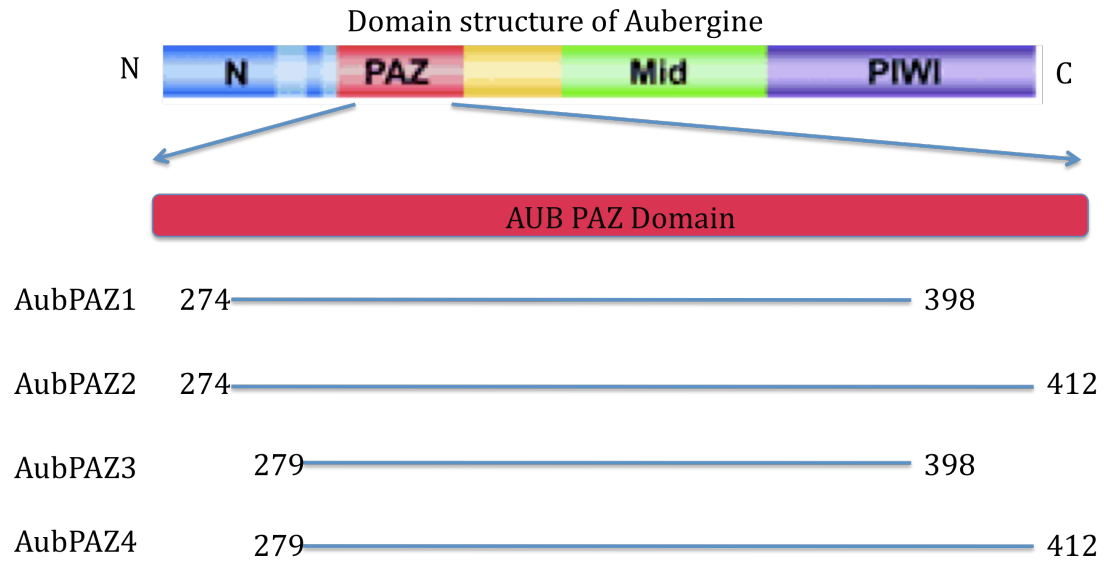



**Figure 39: Aubergine PAZ domain constructs summary. The number represents the amino acid residue number from the N- terminus of the full length Aubergine protein**

### 3.3.3. Expression and Solubility testing

The expression tests results for AubPAZ constructs are summarised in **Table 11**. All four constructs showed comparable soluble expression in *E. coli* BL21(DE3) pLysS system (Stratagene), when protein expression was induced by adding 0.1 mM IPTG at 18°C with overnight expression. AubPAZ3$_{(279\text{-}398)}$ was subsequently chosen for the pursuit of crystallization as it was the shortest construct with less flexible regions that might hinder crystallization.

| Construct | Vector Used | Tag | Expression Temperature | Solubility |
|---|---|---|---|---|
| AubPAZ1$_{(274\text{-}398)}$ | pTWO-E | His | 18°C | ✔✔ |
| AubPAZ1$_{(274\text{-}398)}$ | pTWO-E | His | 25°C | ✔ |
| AubPAZ1$_{(274\text{-}398)}$ | pTWO-E | His | 37°C | ✘ |
| AubPAZ2$_{(274\text{-}398)}$ | pTWO-E | His | 18°C | ✔✔ |
| AubPAZ2$_{(274\text{-}412)}$ | pTWO-E | His | 25°C | ✔ |
| AubPAZ2$_{(274\text{-}412)}$ | pTWO-E | His | 37°C | ✘ |
| AubPAZ3$_{(279\text{-}398)}$ | pTWO-E | His | 18°C | ✔✔ |
| AubPAZ3$_{(279\text{-}398)}$ | pTWO-E | His | 25°C | ✔ |
| AubPAZ3$_{(279\text{-}398)}$ | pTWO-E | His | 37°C | ✘ |
| AubPAZ4$_{(279\text{-}412)}$ | pTWO-E | His | 18°C | ✔✔ |
| AubPAZ4$_{(279\text{-}412)}$ | pTWO-E | His | 25°C | ✔ |
| AubPAZ4$_{(274\text{-}412)}$ | pTWO-E | His | 37°C | ✘ |

**Table 11: Aubergine PAZ domain constructs expression and solubility testing**

### 3.3.4. Purification of AubPAZ3$_{(279-398)}$

The AubPAZ3$_{(279-398)}$ purification scheme was developed de novo with a final yield around 1.5 mg per 6L prep. It was purified in the sequence of Ni-NTA affinity Column, 3C cleavage overnight, GST-gravity flow column to remove 3C protease, second Ni-NTA column to remove the His- tag and the contaminants, and finally S75 size exclusion gel filtration column (described in details in Chapter 2). The purity of the protein was confirmed by running sample fractions on SDS-PAGE gels. The symmetrical elution peak at the elution volume of 197 ml of AubPAZ3$_{(279-398)}$ on the gel filtration column and the single band of the correlated elution fractions on the SDS-PAGE gel indicated that AubPAZ3$_{(279-398)}$ was a stable construct as shown in **Figure 40**.

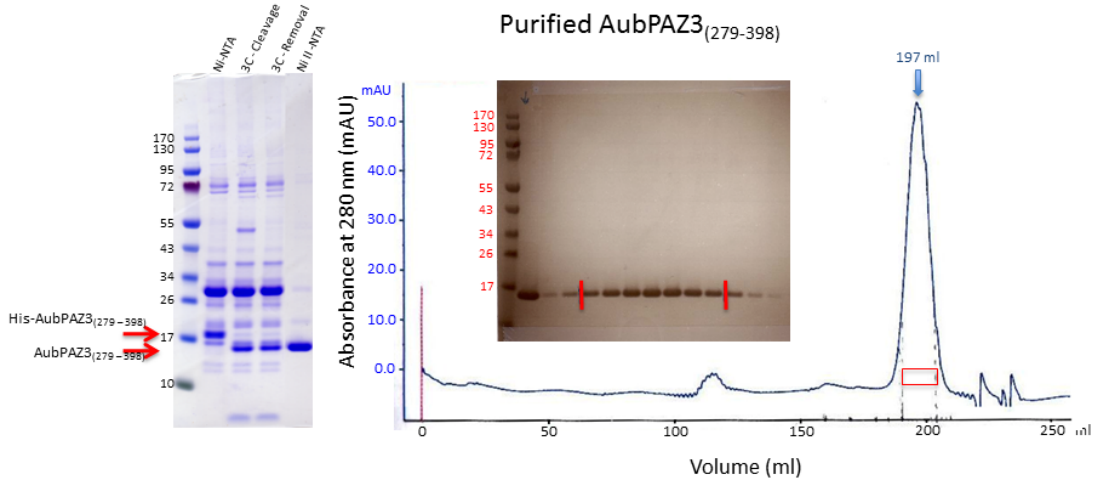



**Figure 40: AubPAZ3$_{(279-398)}$ Purification scheme. Left panel shows the SDS-PAGE gel of the first four steps of the purification. The red – arrow indicating the band corresponding to AubPAZ3$_{(279-398)}$. The right panel shows the Gel filtration trace with SDS -PAGE gel of the eluted fractions. Molecular markers are shown in kilo Daltons**

The identity of AubPAZ3$_{(279-398)}$ was confirmed by N-terminal sequencing and Mass spectrometry shown in **Figure 41**.

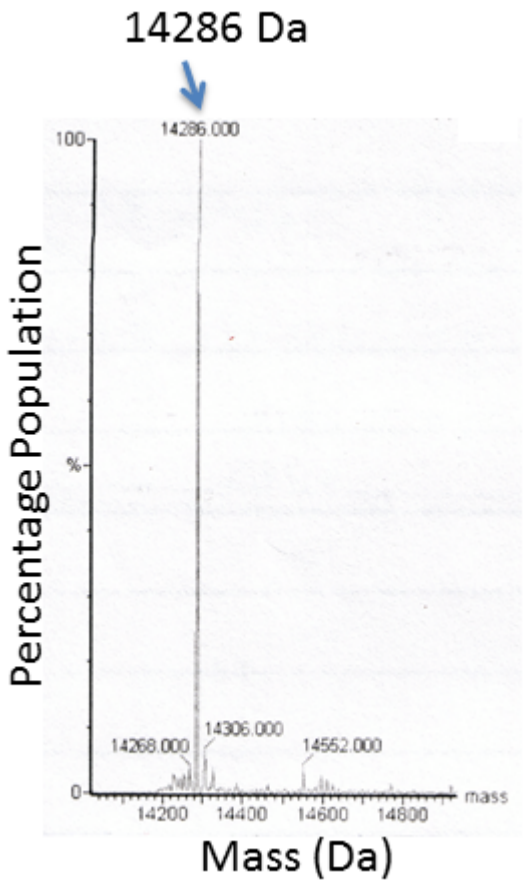



**Figure 41: Mass spectrometry results confirms the molecular weight of the sample corresponds to that of AubPAZ3$_{(279-398)}$**

### 3.3.5. Crystallisation of AubPAZ3$_{(279-398)}$

AubPAZ3$_{(279-398)}$ was concentrated to 5.9 mg/ml using VIVA Spin with 5 kDa molecular weight cut off and stored in 5 mM Hepes pH7.5, 300 mM NaCl, 2 mM $MgCl_2$, 5 mM DTT storage buffer. During the concentration process, it appeared that if a sample was spun longer than 10 minutes or at a rate higher than 4000 RPM, it started to precipitate indicating that the protein became aggregated and insoluble at higher protein concentrations. In addition the concentrated protein would precipitate after thawing on ice at a concentration higher than 6mg/ml. To

stabilise the protein, a minimum of 300 mM NaCl was required in the storage buffer to prevent the protein from precipitating.

#### 3.3.5.1. AubPAZ3$_{(279-398)}$/ RNA1 complex formation

Equal volumes of AubPAZ3$_{(279-398)}$ and RNA1 (AACCUAC-Met) (Table 4 in Chapter 2) with molar ratio of 1:2 were incubated on ice for one hour prior to setting up crystallisation trays. Two final concentrations of AubPAZ3$_{(279-398)}$ were used: 2.36 mg/ml and 1.5 mg/ml. The choice of concentration used for AubPAZ3$_{(279-398)}$ in crystallisation trials was based on the stability of the protein and on the concentration at which apparent RNA binding was observed in the EMSA Gel Shift assays described in section 3.5.3.2.

#### 3.3.5.2. Initial Screening

Initial 96-well and 24-well hanging drop screens were set up for both Apo and AubPAZ3$_{(279-398)}$/RNA1 complex at 4°C and 20°C. Two initial hits for AubPAZ3$_{(279-398)}$/RNA1 complex where AubPAZ3$_{(279-398)}$ concentration was 2.36 mg/ml were obtained from Natrix Screen at 20°C. Crystal clusters appeared in condition 2 (50 mM MES pH5.6, 10 mM MgAcetate, 2.5 M Ammonium Sulphate) overnight, which is shown in the left image in **Figure 42**. Small single crystals also appeared in condition 18 (50 mM Sodium Calcodylate pH6.5, 10 mM MgAcetate, 1.3M Lithium Sulphate) after 5 days which is shown in the right image in **Figure 42**. Both crystals, upon testing, gave very weak diffraction patterns, thus further optimisation of the crystals was required.



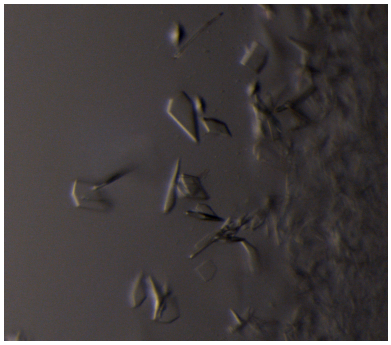


**Figure 42 Left: AubPAZ3$_{(279-398)}$/ RNA1 complex Crystals clusters in Natrix Screen condition 2 20°C overnight. Right: AubPAZ3$_{(279-398)}$/ RNA1 complex crystal in Natrix screen condition 18. Day 5 at 20°C**

#### 3.3.5.3. Optimisation

Unable to reproduce crystals from condition 18 (50 mM Sodium Calcodylate pH6.5, 10 mM MgAcetate, 1.3M Lithium Sulphate), most effort was focused on optimising crystals from condition 2 (50 mM MES pH5.6, 10 mM MgAcetate, 2.5M Ammonium Sulphate). Single crystals were obtained in 50 mM MES pH5.6 and pH6.1, 10 mM MgAcetate, 2.2M Ammonium Sulphate conditions and the diffraction quality was further improved by supplementing glycerol to the condition, as glycerol prevented excessive nucleation of the crystals. Upon testing over 30 crystals by collecting data at a synchrotron, it appeared that the higher concentration of glycerol, the longer it took for the crystals to grow, the better quality data set it produced. The best crystals were obtained in the final condition of 50 mM MES pH6.1, 10 mM MgAcetate, 2.2M Ammonium Sulphate and 16% Glycerol. Crystals appeared 1 month after setting up the screen as shown in **Figure 43**.



**Figure 43: Crystals in 50mM MES pH6.1, 10mM MgAcetate, 2.2M Ammonium Sulphate and 16% Glycerol.**

### 3.3.6. Crystal harvesting, data collection and crystal structure determination

#### 3.3.6.1. Crystal harvesting

Single crystals were cryoprotected by soaking in the cryoprotectant immediately before being flash frozen at 100K using liquid nitrogen. Crystals were soaked stepwise in increasing concentrations of cryoprotectant (8%, 10%, 12%, 14% and 16% Glycerol content substituting the water content in mother liquor).

#### 3.3.6.2. Data Collection

Data were collected at Diamond I04-1 and ESRF ID14-4. Datasets were indexed and integrated using iMosflm (Battye et al) or Xia2 (Winter, 2010) programs, and scaled and merged using SCALA (1994) from the CCP4 suite or Xia2 (Winter, 2010).

#### 3.3.6.3. SelenoMethionine AubPAZ3$_{(279-398)}$ expression, purification and crystallisation

As there were no known homologue structures or domains with high sequence homology to perform straight forward molecular replacement, effort was made towards obtaining SelenoMethionine crystals for phasing of the structure.

AubPAZ3$_{(279-398)}$ DNA was transformed into B834 (DE3) plysS competent cells which were methionine auxotrophs strain to allow the protein of interest to be labelled with added selenomethionine. A single colony was inoculated into 50 ml LB overnight and spun down at 3000 RPM. The pellet was resuspended in 25 ml selenomethione medium (Molecular Dimensions) as a starter culture. Protein expression (12 Litres) was induced by adding 0.1 mM IPTG at 18°C overnight and protein purification was carried out in the same way as wild type AubPAZ3$_{(279-398)}$ as described in Section 2.8. However the yield was only 5% of that of the wild type protein providing insufficient material for crystallisation.

#### 3.3.6.4. Structure Determination by Molecular Replacement

As described in Section 3.3.6.3, the selenomethione AubPAZ3$_{(279-398)}$ had a very low yield and was insufficient for crystallisation. Instead a molecular replacement program, Phaser (Read, 2001), was used to obtain the phase information in order to solve the structure.

#### 3.3.6.5. Structure solutions

An ensemble of three known PAZ domain structures including *Giardia intestinalis* Dicer (20% sequence identity), *Drosophila* Ago2 PAZ domain (9% sequence identity) and human Ago1 PAZ domain (11% sequence identity) was

used as a search model for molecular replacement. Quite astonishingly a single solution was obtained. Post phasing density modification and phase improvement were carried out in automated program Pirate and Parrot (Cowtan, 2000) of the CCP4 suite. Initial model building and refinement was then carried out using Buccaneer (Cowtan, 2006)/ Refmac program (Murshudov et al, 1997), followed by further model building by the ARP/wARP (Perrakis et al, 2001) and Flex – Warp (Cohen et al, 2008) program in CCP4 suite. Structure solutions and automated model building were carried out by James Parker, Martin Nobel and Muhan Wang.

### 3.3.7. Manual model building and refinement

Automated model building produced a preliminary model. Importantly, plausible densiy for RNA was present, associated with the PAZ domain.

Subsequently, several cycles of manual model building in COOT (Emsley & Cowtan, 2004) and refinement in Refmac and PHENIX.REFINE (Adams et al, 2002) were carried out to build the RNA and further improve the model. Initially, only two RNA nucleotides from the 3'-end could be built into the density map. Simulated annealing in PHENIX.REFINE was then employed for refinement and it improved both the quality of electron density map and the R-factor value. TLS groups were then generated using the PHENIX_Find_TLS_Group tool. Subsequent refinements then included TLS refinement with individual coordinates and isotropic Atomic Displacement Parameter (ADP) refinement. Non-Crystallographic Symmetry (NCS) restrains were also used during the refinement. Additional amino acids and RNA nucleotides were manually built

into the structure in COOT after further electron density became visible with the refinement. Several rounds of manual building and refinement were carried out and the quality of the electron density map was gradually improved. Water molecules were then added to the structure using the PHENIX_Water_Picking tool and manually checked in COOT to remove the unlikely ones. The addition of water molecules as well as the metal and salt ions had also improved the quality of the electron density map, which allowed further RNA nucleotides to be built into the structure. Finally an optimisation of weight between the x-ray target and coordinate/ADP restraints was run in PHENIX.REFINE to further lower the R-values.

The final model with a crystallographic R-factor of 0.1895 and $R_{free}$ of 0.2078 comprised of two polypeptide chains (A and B), two RNA molecules (C and D), 81 water molecules, four $Mg^{2+}$ ions and five $SO_4^{2-}$ ions. **Table 12** summarises the data collection and structural refinement statistics. The structure will be discussed in detail in Section 3.4.

| Data Collection | |
|---|---|
| Beamline | Diamond I04-1 |
| Wavelength | 0.92453 |
| Space Group | C2 |
| Unit Cell dimensions a, b, c (Å) | 126.64, 32.88, 86.54 |
| Unit Cell angles $\alpha$, $\beta$, $\lambda$ (°) | 90, 119.97, 90 |
| Resolution (Å) | 27.66 – 1.98 |
| Completeness (%) (Highest Resolution Shell | 98.6 (98.6) |
| Multiplicity | 3.6 (3.4) |
| I/$\sigma$I | 17.5 (2.5) |
| Rmerge (%) | 3.3 (42.8) |
| Number of reflections | 78771 |
| Number of unique reflections | 21834 |
| Refinement | |
| Rwork/Rfree (%) | 18.95 / 20.78 |
| Rmsd from ideal | |
| Bond length (Å) | 0.023 |
| Bond angles (°) | 2.19 |
| Ramachandran plot | |
| Favoured (%) | 97.9% |
| Allowed (%) | 99.5% |

**Table 12: Crystallographic data and refinement statistics for AubPAZ3$_{(279-398)}$/RNA1 complex**

### 3.3.8. Discussion

Eukaryotic Argonaute proteins are known to be notoriously difficult to express and purify even for isolated domains for structure studies, as they are likely to exist with multiple binding partners in *vivo* and there are no existing full length eukaryotic Argonaute protein structures. A significant amount of effort was made to produce crystallisable quality and quantities of the PAZ domain (AubPAZ3$_{(279\text{-}398)}$)of Aubergine protein, a *Drosophila* Piwi subfamily of Argonaute protein in order to produce the first structure of a Piwi subfamily PAZ domain and examine the nature of the recognition of a methylated 3'-piRNA end. Although the expression and purification of AubPAZ3$_{(279\text{-}398)}$ were relatively straight forward, (since the protein still binds to the Ni – NTA column with a lower affinity after the His-tag was cleaved off, providing a quick way to remove the majority of impurities during purification); the stability of the AubPAZ3$_{(279\text{-}398)}$ protein was much more problematical. As indicated in section 3.3.5, AubPAZ3$_{(279\text{-}398)}$ started to precipitate and became insoluble if the concentration exceeded 6 mg/ml (250 μM) or in less than 300 mM NaCl containing buffer. This proved to be challenging for crystallisation as it was particularly difficult to avoid AubPAZ3$_{(279\text{-}398)}$ from precipitating during the concentration step prior to setting up crystallisation trials. In addition, high salt concentration could potentially disrupt the formation of an RNA- protein complex. The instability of the native AubPAZ3$_{(279\text{-}398)}$ protein could also have potentially contributed to low level expression and instability of SelenoMethione AubPAZ3$_{(279\text{-}398)}$.

Still, after much effort, initial hits of crystals for complex of AubPAZ3$_{(279-398)}$ and single stranded RNA with 2'-O-methyl group at the 3' end were obtained using the optimised concentration of protein and RNA mix derived from binding assays. Various optimising and cryo-protecting methods were employed and tested to improve the quality of the crystals and data set. The highest resolution and quality data set was collected finally on crystals that grew very slowly in a high glycerol supplement. During the course of crystallisation optimisation, it was demonstrated that the quality of the dataset was proportional to the amount of glycerol in the precipitant.

The structure was solved by molecular replacement after the expression and purification of SelenoMethionine AubPAZ3$_{(279-398)}$ were proven to be not viable. The structure and its implication are described further in the next Section.

## 3.4. Structure Analysis of AubPAZ3$_{(279-398)}$ / piRNA complex

### 3.4.1. Introduction

Crystals of AubPAZ3$_{(279-398)}$ in complex with 7-mer RNA (5'-AACCUAmC-3'; i.e. A1 to mC7) 2'-O-methylated at the 3'-end were obtained through vapour diffusion and diffracted to 1.98 Å. After initial molecular replacement, the majority of the protein polypeptide chains were built through automatic model building programs. The final model of the AubPAZ3$_{(279-398)}$/7mer RNA with 2'-O-methylation at 3'- end complex (space group C2) was completed after cycles of manual building (including the RNA molecules) and refinement.

### 3.4.2. Aubergine Argonaute PAZ domain – piRNA structure

#### 3.4.2.1. Overall structure of AubPAZ3$_{(279-398)}$ with single stranded 2'-O-methyl RNA

The Aubergine PAZ domain consists of a central β barrel core, which is formed by five β strands (β1, β2, β5, β6 and β8) and a β turn (β7). At one side of the core there is an αβ motif consisting of an α helix and a β hairpin (α3, β3 and β4). On the other side of the core there are two α helices. The overall schematic representation of the topology of the Aubergine PAZ domain is shown in **Figure 44**.

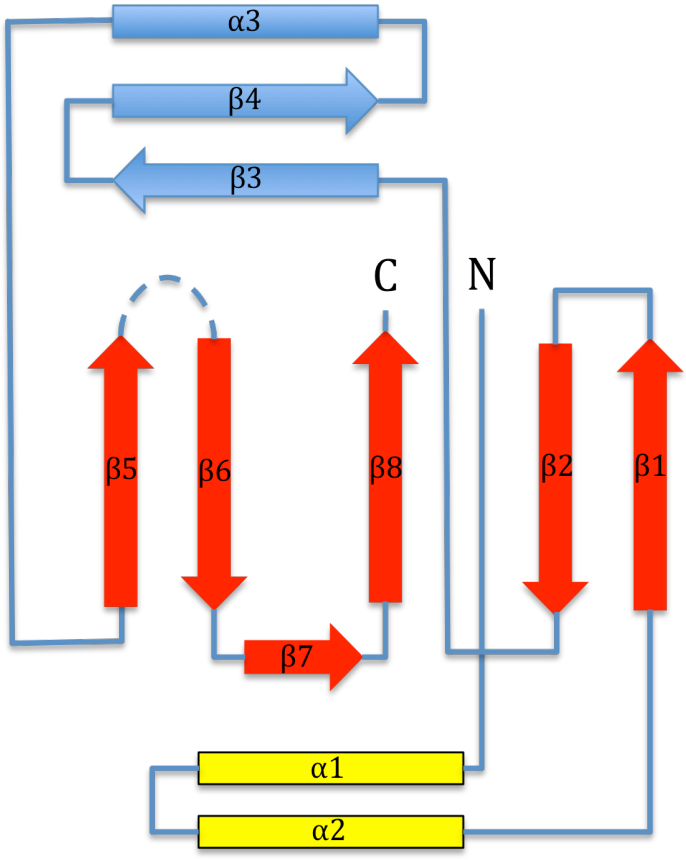



**Figure 44: Schematic representation of the topology of the Aubergine PAZ domain. The beta barrel core (in red) is formed by five beta strands and an alpha-beta turn. The core is flanked by two alpha helices at one side (in yellow) and an alpha-beta motif on the other side (in blue).**

**Figure 45** shows the overall structure of the Aubergine PAZ domain in complex with a 7mer 2'-O-methyl RNA. As shown in **Figure 45** the αβ motif (α3, β3 and β4) form one side of the wall of the RNA binding pocket; β2, β3 and β4 strands from the central β barrel core form the other side of the wall of the RNA binding pocket.

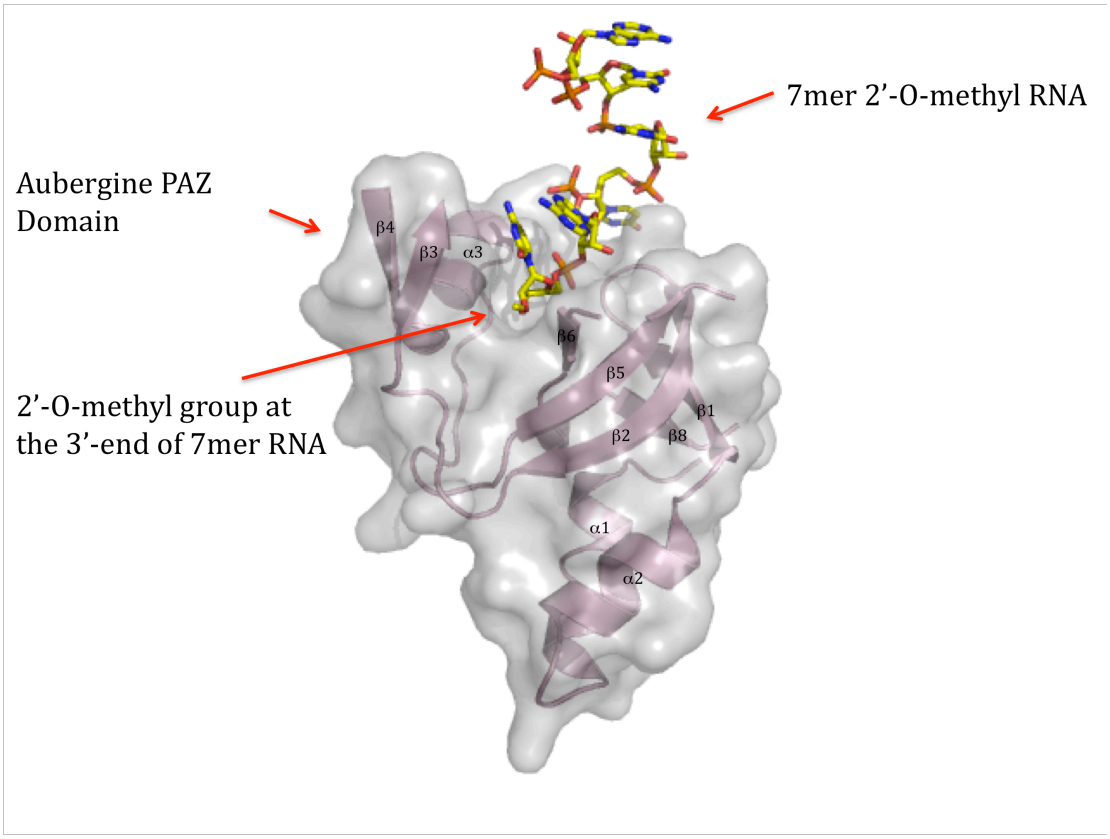



**Figure 45: Structure of Aubergine PAZ domain in complex with the 7mer 2'-O-methyl RNA (5'-AACCUAmC-3'). Figure was generated with Pymol, as all structure illustrations in this thesis.**

The Aubergine PAZ domain RNA binding pocket is lined by a number of hydrophobic and aromatic residues (Tyr 313, Tyr318, Phe333, Tyr342, Tyr345, Tyr346, Tyr350, Ile379, Met380 and Ile381). Aromatic residues are commonly found at nucleic acid binding sites. These residues are conserved across the Piwi subfamily of Argonaute proteins shown in the sequence alignment **Figure 46**,

suggesting that the PAZ domains of the Piwi subfamily of Argonaute proteins employ similar mechanisms in binding the 3' – end of piRNAs.

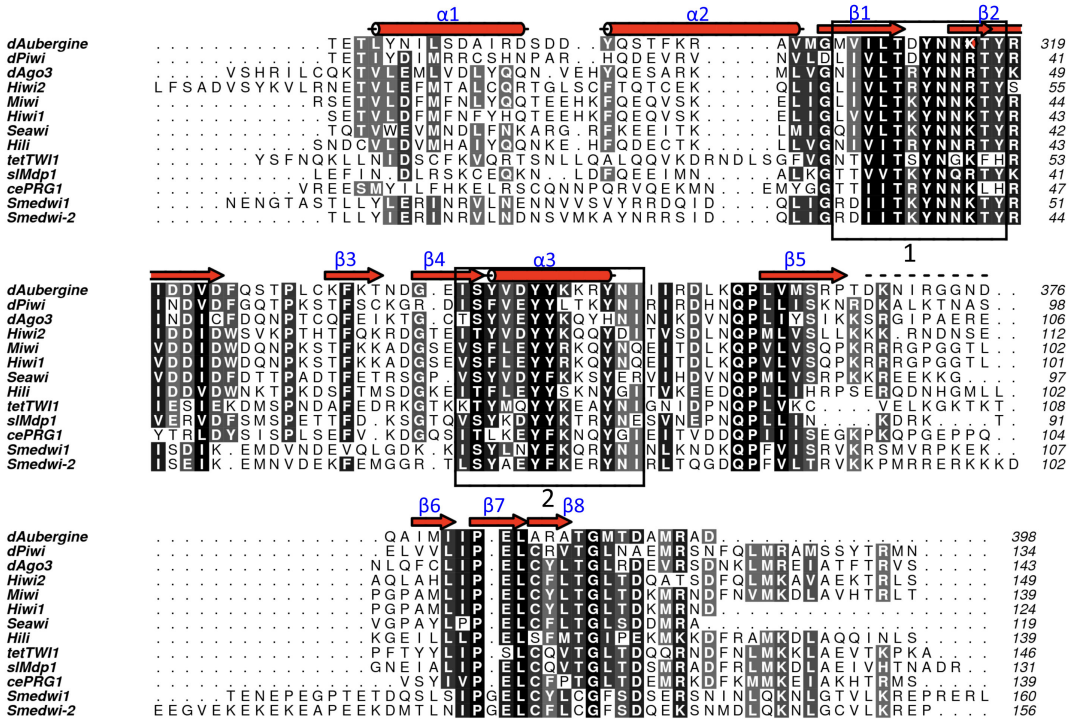



**Figure 46: Sequence alignment of PAZ domain for the Piwi subfamily of Argonaute proteins. The secondary structures (helices and beta strands) of Aubergine PAZ domain are shown across the top. Box 1 shows the conserved residues that are specific to the Piwi subfamily of Argonaute proteins; Box 2 shows the conserved residues that are shared between both the Ago and Piwi subfamilies of PAZ domains.**

As shown in **Figure 47**, the 3'–end of the single stranded RNA is inserted deeply into the RNA binding cleft, the first two nucleotides especially making multiple contacts with the residues from the α3 helix, β3 (of the αβ motif) and β6 strand (of the barrel) in a sequence non-specific manner. The majority of the interactions are formed between the sugar-phosphate backbone and the ribose rings with amino acid residues lining the binding cleft. This is consistent with the ability of Argonaute proteins to bind to short RNAs of all sequences.

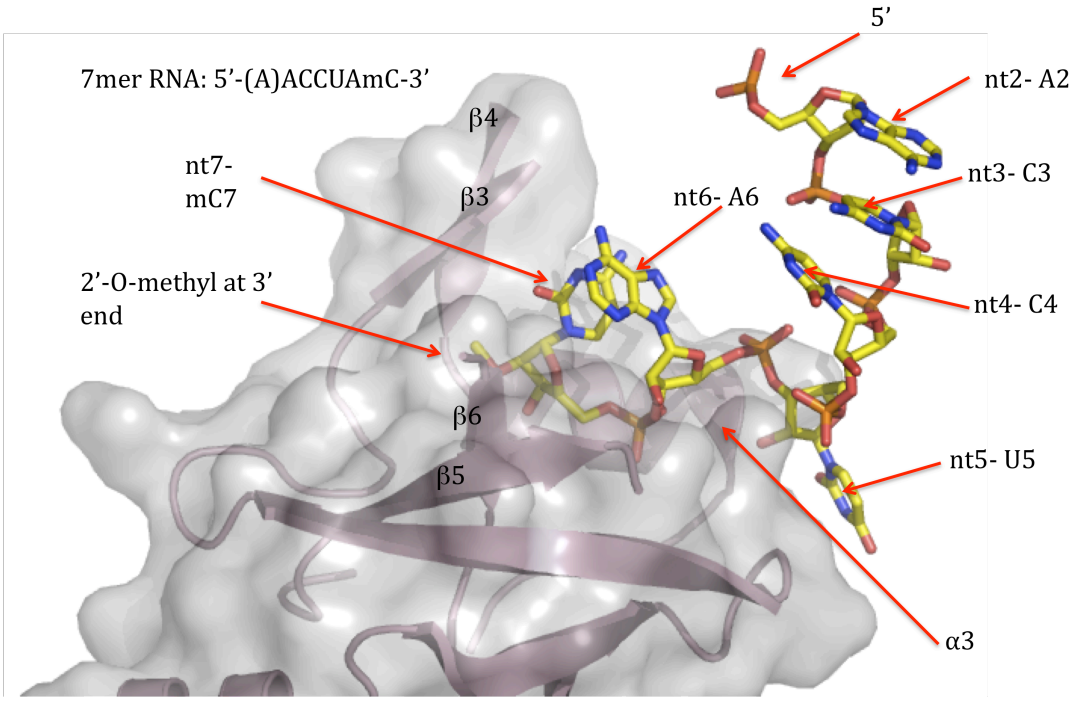



**Figure 47: The overall structure of binding pocket of Aubergine PAZ domain. 7mer RNA with 2'-O-methylation at the 3'-end bound to the Aubergine PAZ domain.**

The 2’ – O- methyl group at the 3’ – end of the piRNA is positioned towards a hydrophobic groove in the RNA binding pocket formed by the β hairpin (β3 and β4 strands) at one side and the β5, β6 strands (of the barrel) on the other. A number of hydrophobic residues including Phe333, Ile379, Met380, Ile381 provide a hydrophobic environment shown in **Figure 48**.

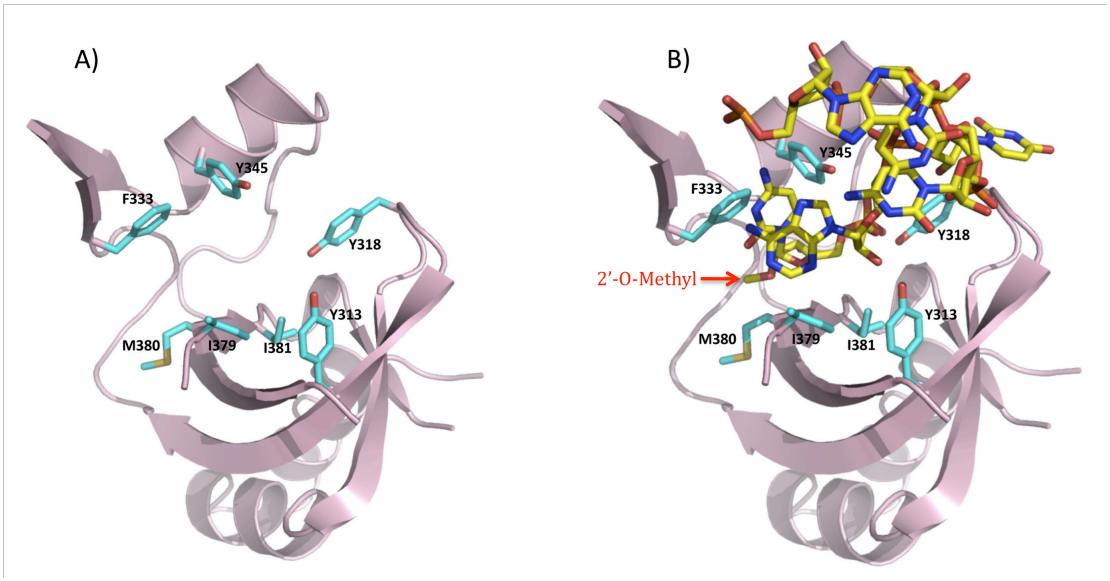



**Figure 48: Top down view of the Aubergine PAZ domain RNA binding pocket. A) (RNA is removed). The hydrophobic and aromatic residues line the RNA binding pocket. B) RNA-bound view of the Aubergine PAZ domain RNA binding cleft. The 3'-end of the piRNA is inserted into a binding pocket that is lined by a number of hydrophobic and aromatic residues.**

The third nucleotide from the 3' – end is positioned at the edge of the binding pocket and makes contacts with the surface of the protein. The electron density map reveals a further three nucleotides towards the 5' - end away from the protein as shown in **Figure 48**. Overall the RNA molecule projects outwards from the PAZ domain with it's 3'- end anchored at the heart of the binding cleft.

#### 3.4.2.2. Two molecules in the asymmetric unit

There are two complexes in the asymmetric unit as shown in **Figure 49**, the discussion of this section refers to the better defined Molecule A.

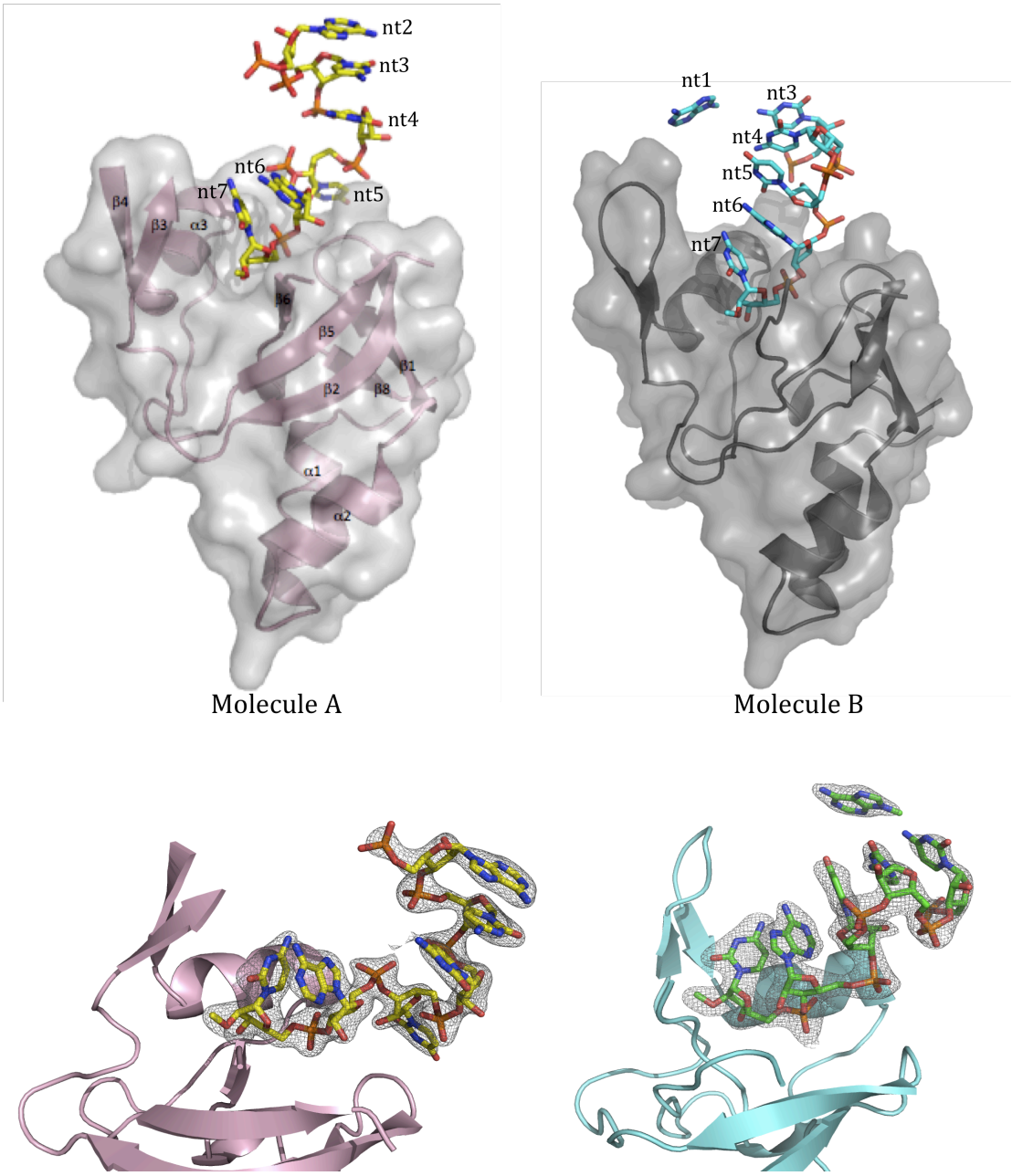



**Figure 49: Two molecules in the asymmetric unit of the crystal structure. The electron density shows the RNA in each of the molecule in the asymmetric unit.**

### 3.4.2.3. Detailed recognition of RNA binding – involvement of conserved residues

**Figure 50** shows the electron density (contoured at 1.5 σ) of the 3'- end of the RNA in the binding pocket of the PAZ domain. There is a clearly defined density for the 2'-O-methyl group at the 3'-end of the RNA.

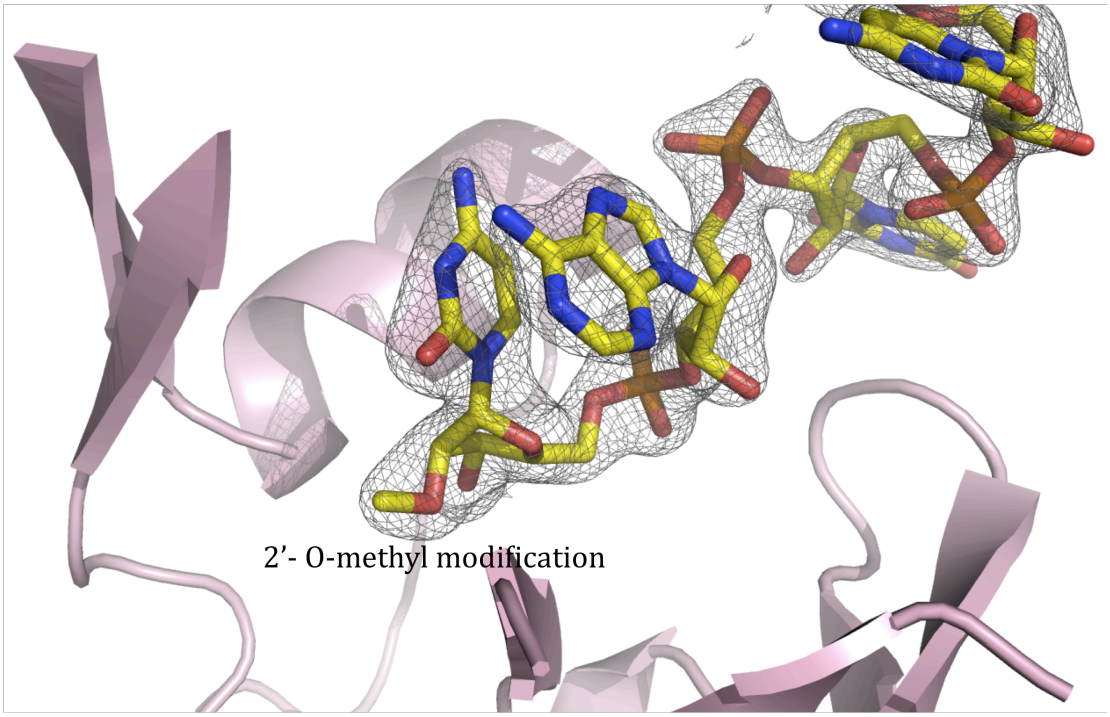



**Figure 50: Electron density at the 3'- end of the RNA (contoured at 1.5 σ).**

The RNA binding pocket in the Aubergine PAZ domain is a “V – Shaped” binding cleft formed by an αβ module (α3 helix, β3- β4 hairpin) at one side and β5-β6 strands on the other. The 3’ end of the piRNA (especially the last two nucleotides A6 and mC7) interact with a number of aromatic and hydrophobic residues that line the binding cleft as shown in **Figure 51**.

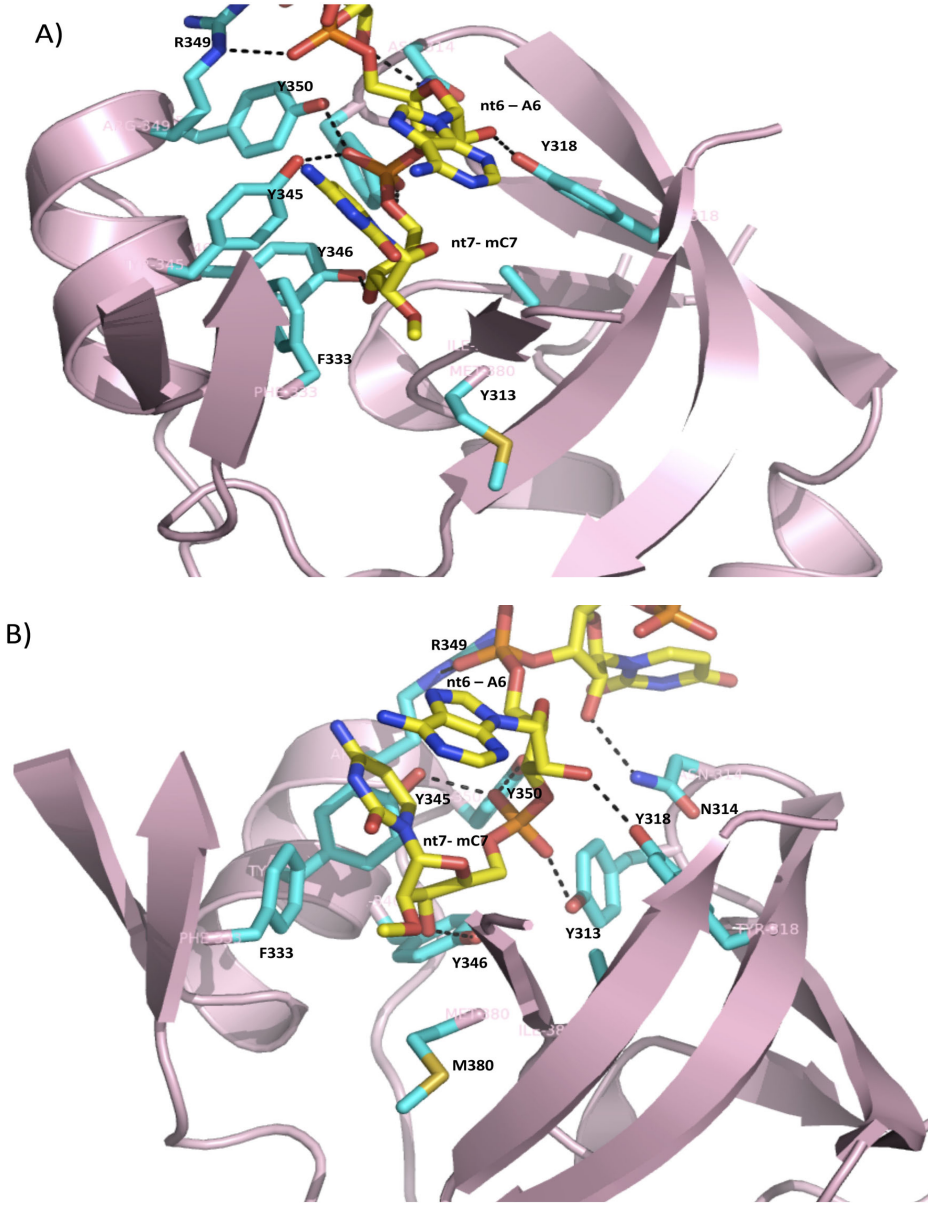



**Figure 51: Detailed recognition of 3'-end of piRNA by Aubergine PAZ domain RNA binding pocket. The dotted black lines are the hydrogen bonds formed between the RNA and the Aubergine PAZ domain residues. A and B show two different orientations of the binding pocket.**

Tyr318 forms a hydrogen bond with the –OH group on the ribose ring on the A6 nucleotide; Tyr345 forms a hydrogen bond with the non-bridging oxygen on the phosphate group on mC7; Phe333, Tyr345 and Ile381 contribute to generating a hydrophobic environment at the heart of the binding cleft. These residues are invariant throughout both PAZ domains of the Piwi and Ago subfamily of Argonaute proteins implying that the general mode of RNA binding is consistent across Argonaute proteins.

A number of interactions were observed between the last nucleotide mC7 and the residues on the binding cleft. Tyr313 and Tyr350 interact with the non-bridging oxygen on the phosphate group of mC7; Tyr346 forms hydrogen bonds with the hydroxyl group on the ribose ring of mC7 and this renders the specific recognition of RNA as the substrate, since the 2'-OH group does not present on DNA. Thus the structure of PAZ domain reflects its binding specificity. These residues are invariant throughout the Piwi subfamily of Argonaute proteins but vary across the Ago subfamily, suggesting that the recognition of the 3' – end of the piRNA in Piwi Argonaute PAZ domains could be subtly different.

Additional intermolecular interactions are observed between nucleotide 5 (U5) and 6 (A6) and the binding cleft residues. Asn314 forms hydrogen bonds with the hydroxyl group on the U5 ribose group; Arg349 contacts non-bridging oxygen on the phosphate group on A6; Tyr342 and Ile379 also contribute to the formation of the hydrophobic binding cleft for accommodating the 2' – O- methyl group at the 3' – end of the piRNA. These residues are also highly conserved within the Piwi subfamily of Argonaute proteins.

The structure of the binding pocket of Aubergine PAZ domain shows conserved RNA substrate recognition. The substrate recognition is RNA specific but non – sequence specific.

### 3.4.2.4. Recognition of 2’-O-methyl group in a hydrophobic pocket

The 2’-O-methyl group at the 3’- end of piRNA (mC7) is inserted deeply into a hydrophobic groove lined with a number of conserved hydrophobic and aromatic residues. The presence of residues such as Phe333, Tyr342, Ile379, Met380 and Ile381 contribute to generate a hydrophobic environment for 2’ – O-methylation modification at the 3’ – end of the piRNA, as demonstrated in **Figure 52**.

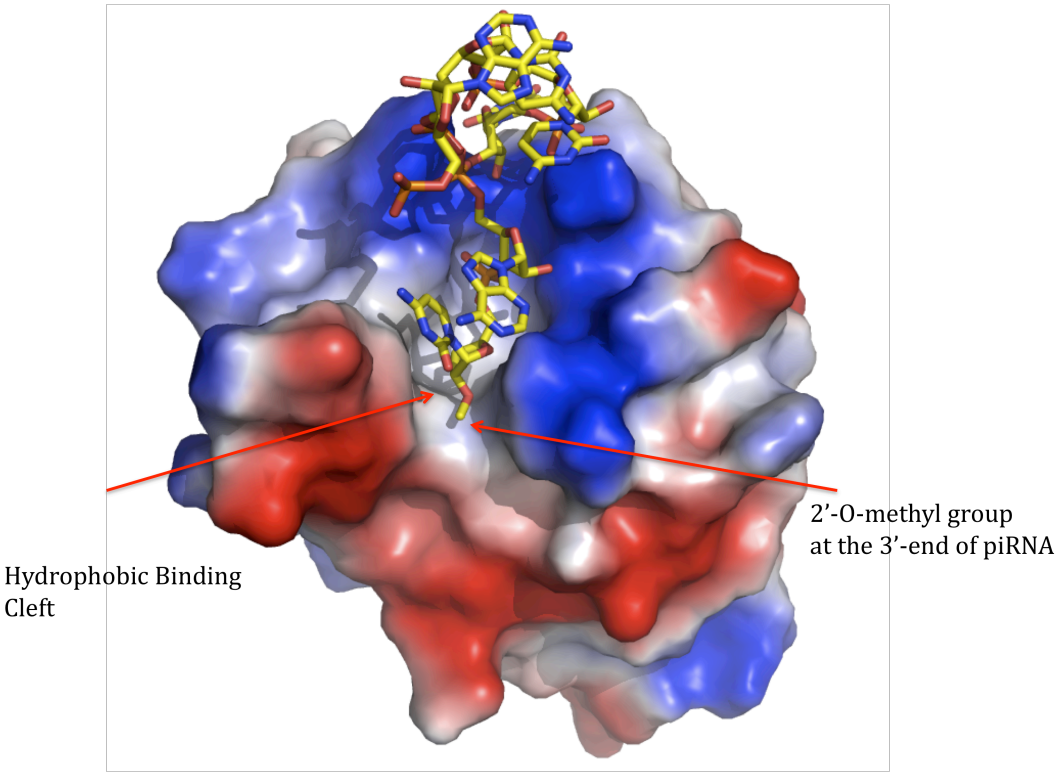



**Figure 52: Charged surface representation of the Aubergine PAZ domain protein bound to the 7mer RNA with 2'-O-methyl modification at the 3'-end. The white colour indicates a non-charged patch. The charged surface is generated by Pymol.**

The surface view of the binding pocket in complex with the 3’ – end of the piRNA in **Figure 52** shows that there was room in the hydrophobic binding cleft to accommodate the presence of the 2’-O-methyl group and there are no steric clashes observed. This mode of recognition is compared to that of the Ago subfamily PAZ domain and will be discussed in detail in Section 3.4.3.

### 3.4.2.5. A 'flipped' nucleotide

One interesting and unexpected feature of the structure was that, in one of the two molecules in the asymmetric unit, the 3rd nucleotide from the 3' – end of the piRNA (U5) is "flipped" out at the opening of the RNA binding pocket, as shown in **Figures 53 and 54**.

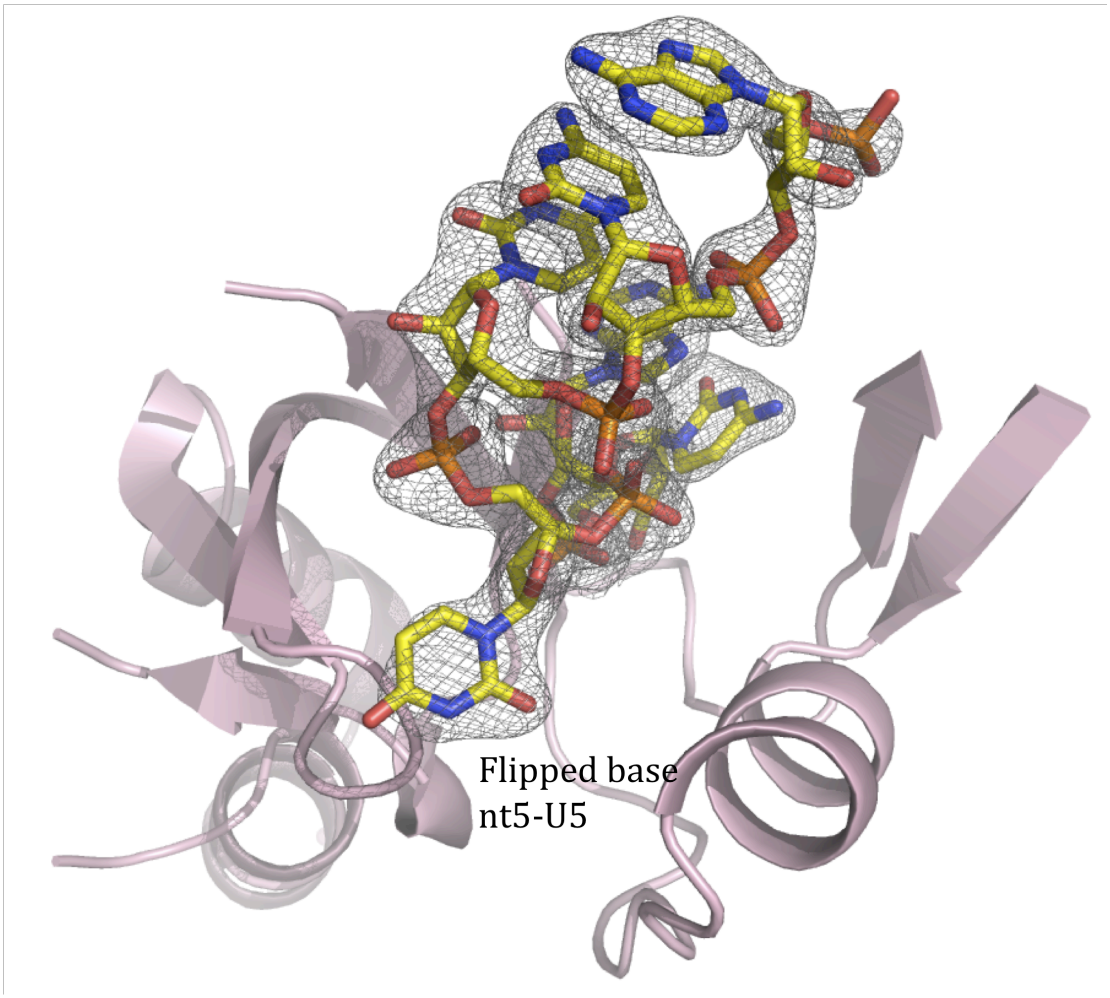



**Figure 53: Electron density for the 'flipped' base nucleotide.**

Instead of following the helical path, the ribose group and the base of U5 are both rotated onto and supported by the surface of the PAZ domain leaving an apparent gap in the stacking of the RNA bases. **Figure 53** shows the electron density for the "flipped" base and **Figure 54** shows the interactions between the "flipped" base nucleotide and the Aubergine PAZ domain.

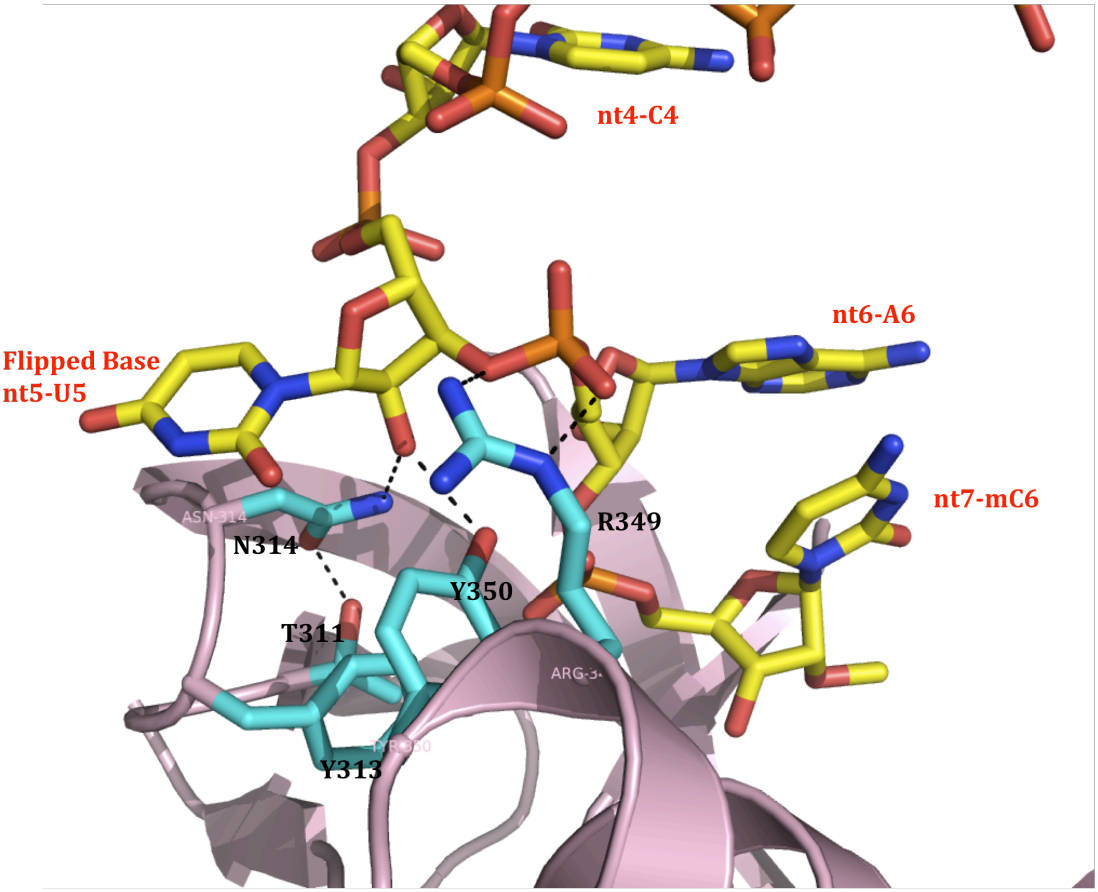



**Figure 54: The 'flipped' nucleotide. The base of nucleotide 5 (U5) is "flipped" and this conformation is stabilised by several hydrogen bonding interactions shown in black dotted lines with a number of conserved residues of Aubergine PAZ domain protein.**

The 2'-OH group on the ribose ring is hydrogen bonding with Asn314, and Tyr350. Arg349 forms additional hydrogen bonds with both the bridging oxygen and the non-bridging oxygen of the phosphate group. The base is further stabilised by interaction with the carbonyl group of Tyr313. Overall the electron density of the U5 nucleotide is very well defined (**Figure 53**) owing to the large number of interactions stabilising this confirmation. The quality of the electron density map confirms the presence of this unusual conformation.

Tyr313, Tyr350 and Asn314 are amongst the most conserved residues across the Piwi subfamily of Argonaute proteins. However, they are varied in the Ago

subfamilies. **Figure 55** shows the conserved residues mapped onto the surface of Aubergine PAZ domain.

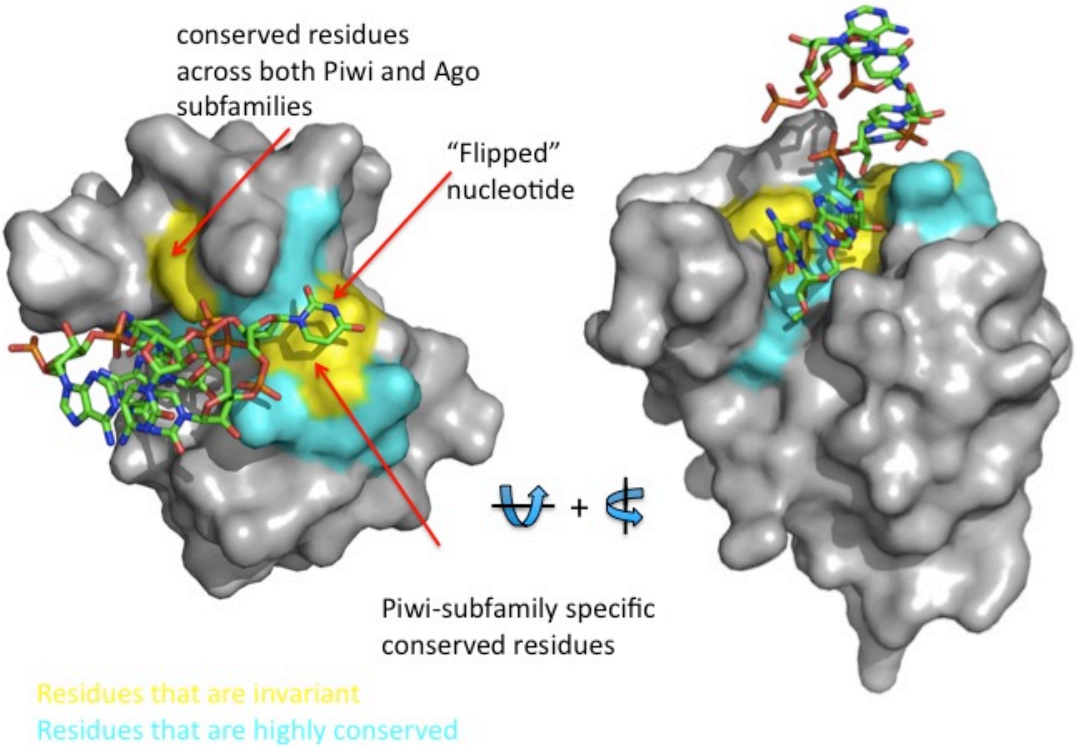



**Figure 55: Sequences conservation mapped onto the surface of Aubergine PAZ domain. The yellow colour shows the most conserved residues that are invariant across the species. The turquoise colour shows highly conserved residues.**

The "flipped" nucleotide localises over some of the most conserved residues as shown in **Figure 55**. Moreover, they are Piwi subfamily specific. This indicates that this conformation maybe important in piRNA recognition.

#### 3.4.2.6. Putative metal binding sites

In the crystal structure of AubPAZ3$_{(279\text{-}398)}$ / 7 mer RNA (5'-AACCUAmC-3'), I assigned four metal ions in the structure. One of these appears to play an important role in coordinating and stabilizing the RNA structure.

This $Mg^{2+}$ ion is located at the surface of the PAZ domain near the opening of the binding pocket as shown in **Figure 56**. It appears to be involved in interacting and stabilizing the sugar phosphate backbone of the RNA molecule.

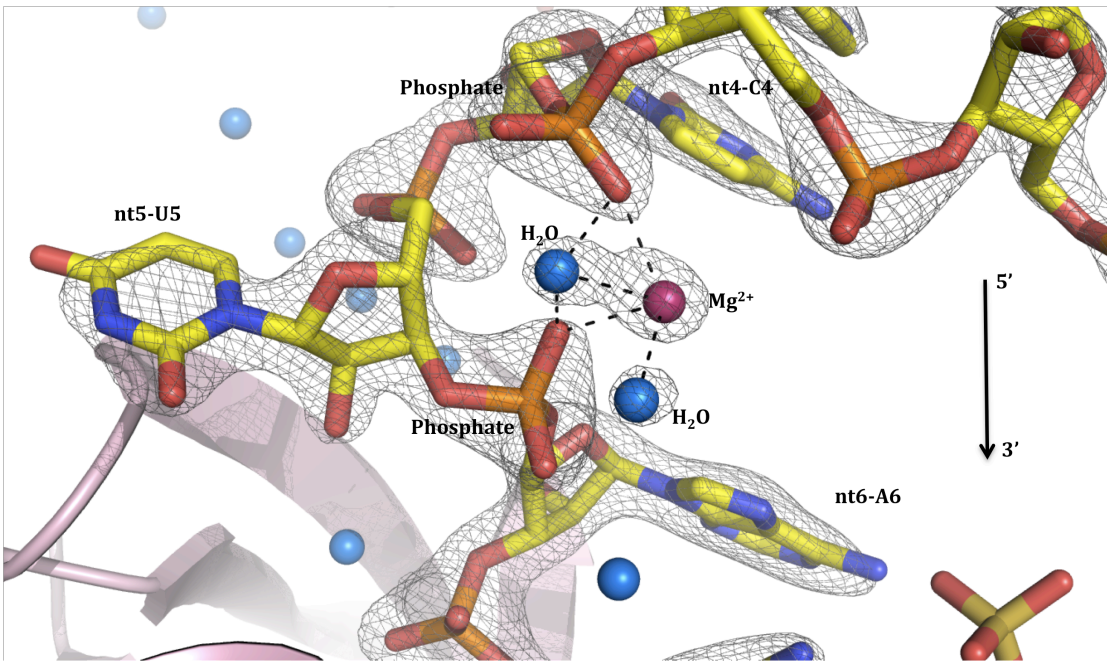



**Figure 56: Electron density representation of putative $Mg^{2+}$ ion stabilising the backbone of the RNA molecule by forming interactions with the phosphate groups and water. Water molecule also forms hydrogen bonds with the phosphate groups.**

As shown in **Figure 56** the $Mg^{2+}$ together with two water molecules coordinate the positioning of the phosphate groups for nucleotide 6 (A6) and nucleotide 4 (C4) simultaneously. The distances between the $Mg^{2+}$ and the phosphate groups are around 2.3 Å, which is the typical distance for metal coordination interactions. As described previously in Section 3.4.2.4, nucleotide 5 (U5) is "flipped" out of the helix and the rotation of the phosphodiester backbone at U5 results in a sharp turn in the backbone and a gap between the two adjacent nucleotides A6 and C4. The $Mg^{2+}$ is positioned such that it stabilizes the positioning of the phosphate groups for A6 and C4. It is located in the apparent gap left by U5 between the nucleotides A6 and C4. Thus this interaction together

with the specific interactions between U5 and the surrounding residues help to stabilize the confirmation of the "flipped" nucleotide

#### 3.4.2.7. Highly coordinated water molecules?

In the crystal structure of AubPAZ3$_{(279\text{-}398)}$ / 7 mer RNA (5'-AACCUAmC-3'), I assigned 81 water molecules. Two (one in each AubPAZ3$_{(279\text{-}398)}$ / 7 mer RNA complex in the asymmetric unit) of them appear to play an important role in coordinating and stabilizing the RNA binding.

The highly coordinated water molecule was located at the heart of the RNA binding cleft as shown in **Figure 57**. It forms hydrogen bonds with Thr311, Asn314, Tyr318 and the non – bridging oxygen on the phosphate group of the terminal nucleotide at the 3'-end (mC7). The co-ordination distances are incompatible with a metal ion.

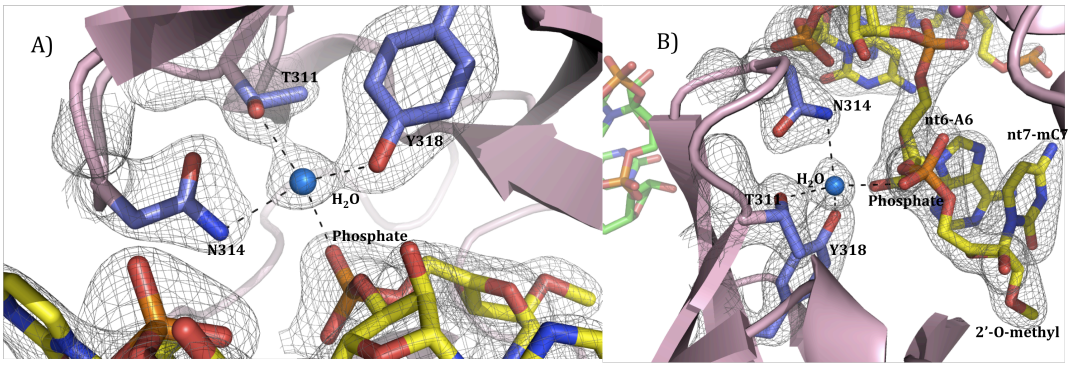



**Figure 57: Highly coordinated water molecule in the RNA binding pocket. A) and B) show two different views of the water molecule.**

As shown in **Figure 57** the water molecule coordinates the three surrounding residues and stabilizes the sugar-phosphate backbone of the mC7 nucleotide in place. This provides another level of RNA 3' – end recognition in addition to the direct interactions between the 3' – end terminal nucleotide and the binding pocket residues.

### 3.4.2.8. The effects of crystal packing

The protein crystal of AubPAZ3$_{(279\text{-}398)}$ / 7 mer RNA (5'-AACCUAmC-3') possesses C 2 space group symmetry. This means that the crystal packing is in part a result of repeating the PAZ-RNA complexes in the asymmetric unit through a two - fold rotation axis. This created a number of intermolecular interactions that potentially contribute to the stabilization of the structure, this is highlighted in **Figure 58**.

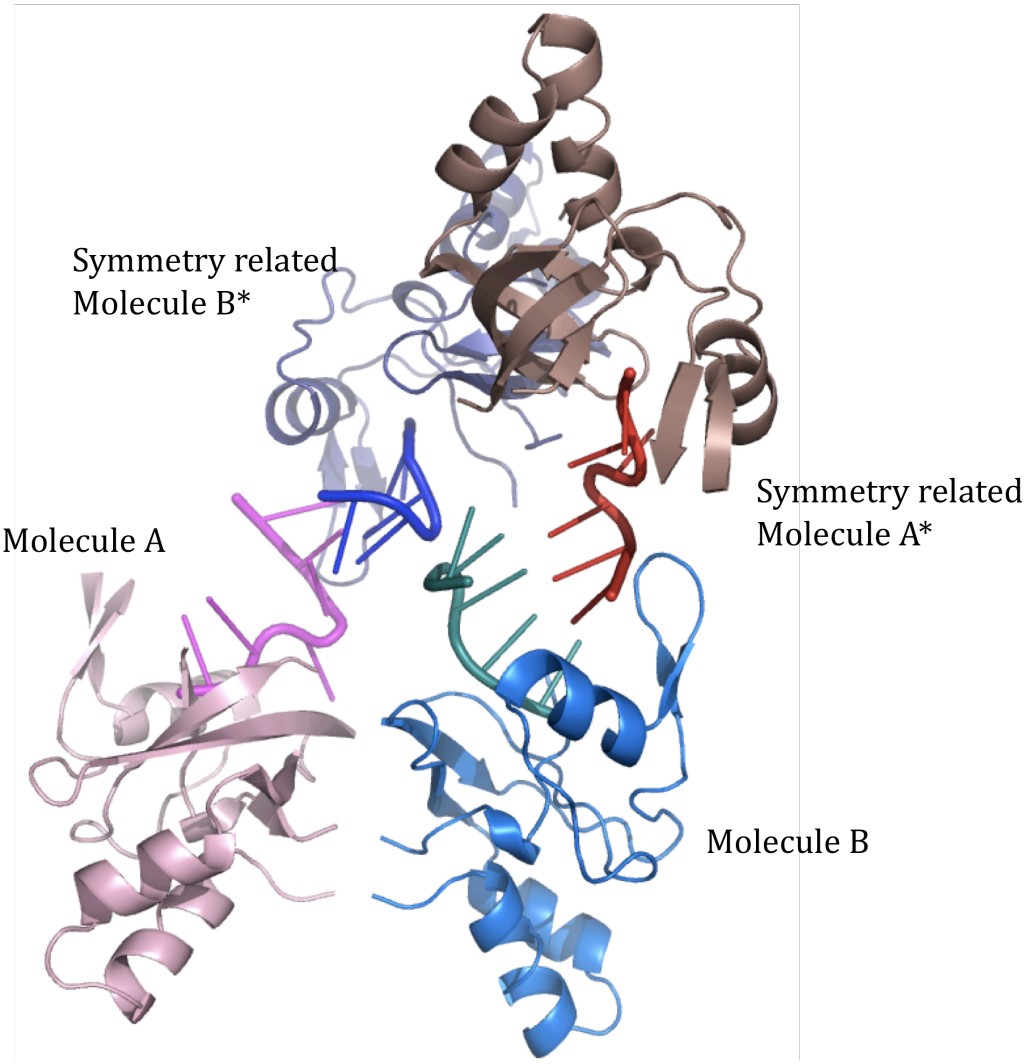



**Figure 58: Molecule A and Molecule B with symmetry related partners.**

**Figure 59A** shows overall RNA interactions through crystal packing in the context of symmetry related molecules.

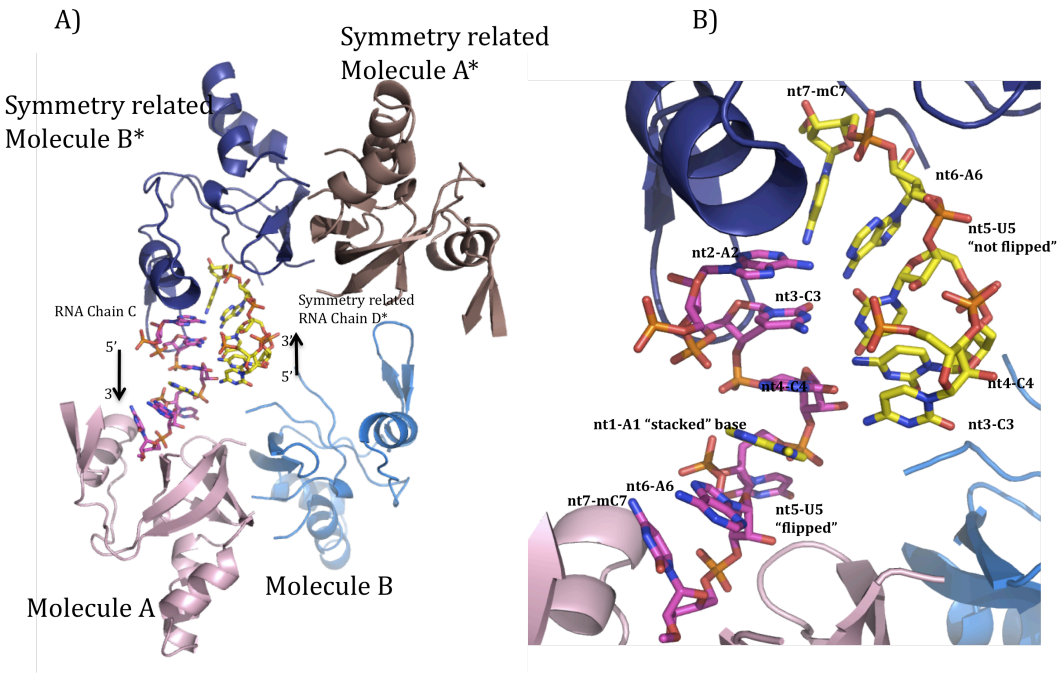



**Figure 59: Symmetry related RNA inter-molecular interactions. B) is the zoomed in view of A).**

As shown in more detail in **Figure 59B**, firstly, the nucleotide 2 – 4 (ACC) from the 5' end of the 7 mer RNA (Chain C) are loosely base paired despite the non-complementary base paring with the nucleotide 4 – 6 (CUA) from the 5' end of the symmetry related 7 mer RNA (Chain D).

Secondly, the Adenine base of the A1 – D* appears to be inserted into an apparent gap in the Chain C between nucleotide 6 (A6 – C) and nucleotide 4 (C4 – C) as shown in **Figure 59**. The gap is formed by the "flipped" nucleotide described in Section 3.4.2.5. The Adenine base is stabilised by the base stacking interaction and it acts as an artificial 5' - anchor for the RNA Chain D* and stabilises the crystal packing. As it is five nucleotides away from the 3' – end nucleotide, this is unlikely to affect the RNA 3' – end binding in the PAZ domain. In *vivo* piRNAs range between 22 – 30 nucleotides in length and the 5' – end would be bound to the MID domain of the Piwi Argonaute protein.

These interactions, mainly at the first 5 nucleotides of the RNA molecules from the 5' – end, contribute to the crystal packing and thus the crystal formation but at the same time do not compromise the specific binding at the 3' end of the RNA. It also explains the existence of the two confirmations shown in **Figure 49** of RNA molecules in the crystal, Molecule A with Chain A and Chain C complex being the more visible complex than that of Molecule B with Chain B and Chain D. Thus Molecule A adopts the main conformation described in this Section.

### 3.4.3. Comparison of the structures of Piwi and Ago Argonaute PAZ domain-RNA complexes

The natural substrates of the Piwi Argonaute PAZ domains contain 2'–O–methyl modification at the 3' – end of the RNA; whereas the Ago Argonaute PAZ domain preferentially bind to RNAs that do not have such modification (i.e. siRNAs and miRNAs). To examine and understand the substrate specificity of the Piwi Argonaute PAZ domain, the structure of Aubergine PAZ domain in complex with a piRNA is compared to the PAZ domain of human Argonaute 1 (hAgo1 – pdb code 1SI2), which is an Ago subfamily Argonaute protein bound to a siRNA.

### 3.4.3.1. Overall structure topology

The AubPAZ3$_{(279\text{-}398)}$ and the hAgo1PAZ show very similar overall structures as shown in **Figure 60**. The two PAZ domains both have a central twisted β barrel core formed by six β strands. Similarly the β barrel core in both structures is flanked by an αβ module (an α helix linked to a β hairpin) at one side and two α helices on the other.

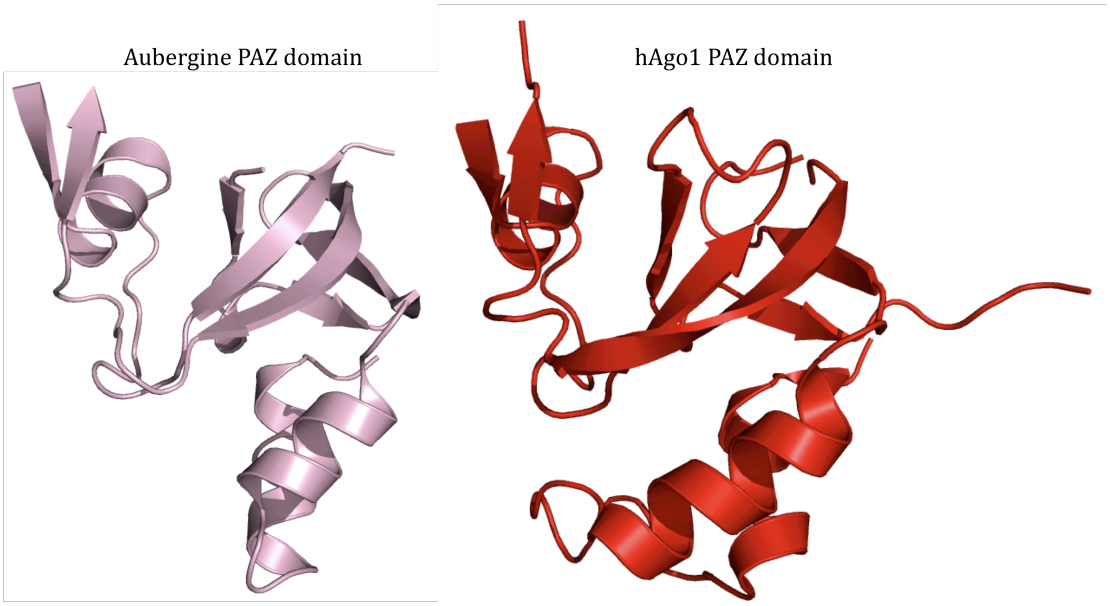



**Figure 60: Overall structural topology of Piwi subfamily (Aubergine) shown in pink and Ago subfamily (hAgo1) PAZ domain shown in red.**

The Piwi and Ago subfamilies of Argonaute PAZ domains only share ~14% of sequence identity and the two superimposed structures (**Figure 61**) have an rmsd (Å) of 1.8 Å. The majority of the structure superimpose well, except that the αβ module in the Aubergine PAZ domain is shifted from that of the hAgo1 PAZ domain creating a slightly larger opening of the binding pocket.

The superimposed structures show that the piRNA bound to the Aubergine PAZ domain is projected outwards from the centre of the binding pocket and only the last three nucleotides at the 3' – end make contacts with the PAZ domain;

whereas the siRNA bound to hAgo1 PAZ domain makes contacts along the phosphate backbone with the surface of the hAgo1 PAZ domain.

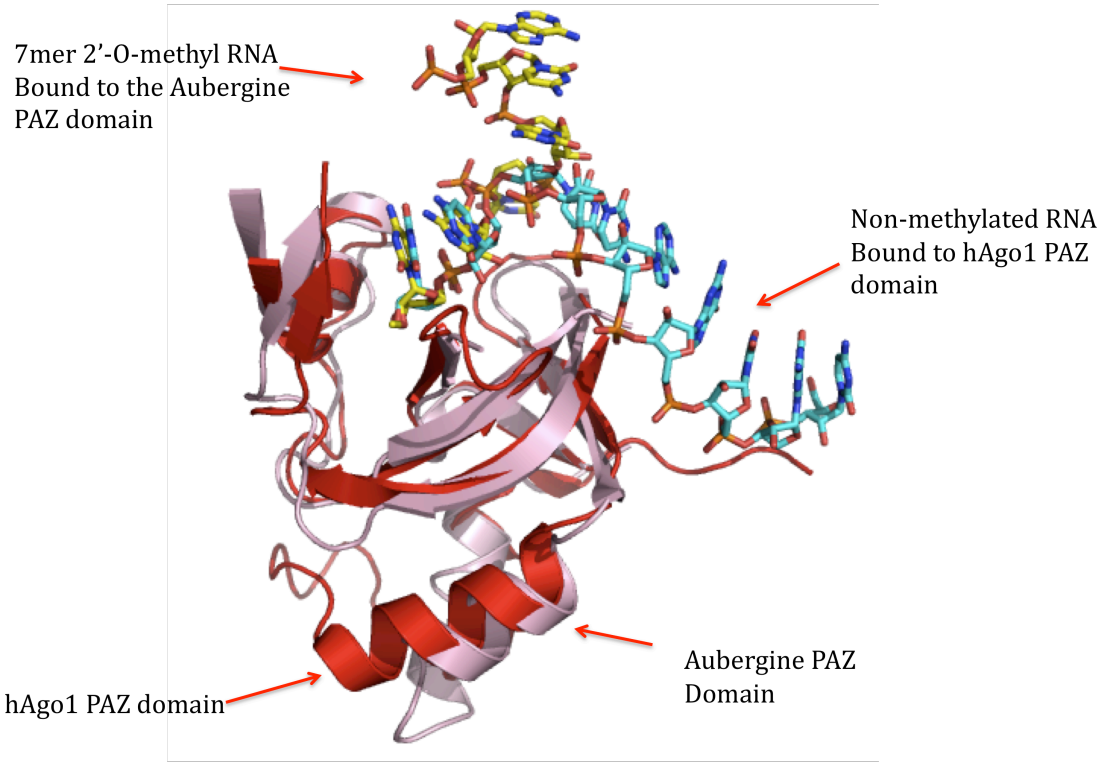



**Figure 61: Superposition of the structures of hAgo1 PAZ domain in red bound to non-methylated small RNA in blue (PDB code 1SI3 (Ma et al, 2004)) with Aubergine PAZ domain in pink bound to a 7mer RNA with 2'-O-methyl modification at the 3'-end in yellow.**

In the crystal structure of hAgo1, it binds to a small RNA duplex that is self-complementary in sequence and thus the structure is a homo-dimer in the crystal. This could explain the trajectory of the small RNA bound to the hAgo1 protein whereas the Aubergine PAZ domain binds to a small RNA, the sequence of which was carefully chosen so that it would remain single-stranded to mimic the natural substrate of Piwi subfamily PAZ domain. Though a 7 – mer RNA is significantly shorter than the natural piRNAs that are usually 22 – 30 nucleotide in length.

### 3.4.3.2. The binding pocket

The Piwi and Ago subfamilies of PAZ domains share a similar topology in the formation of the RNA binding pocket. As shown in the superimposed structure (**Figure 62)**, the binding pockets are formed by the αβ module (α3 helix, β3-β4 hairpin) at one side and the β5-β6 strands on the opposing side. Key residues involved in the RNA binding in the Aubergine PAZ domain are also shown in **Figure 48**.

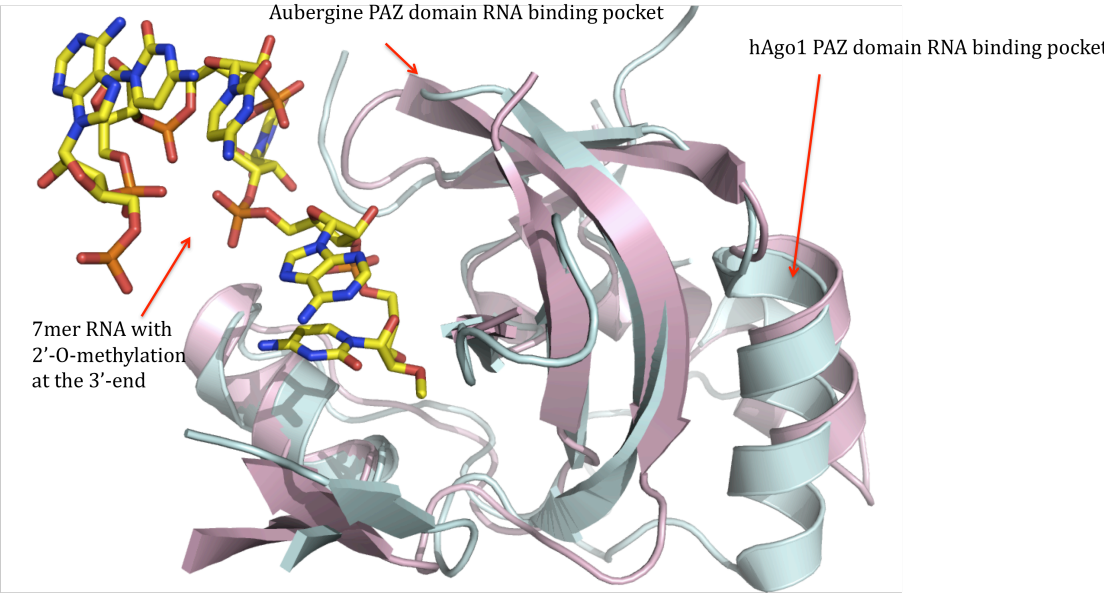



**Figure 62: Superimposed RNA binding pocket of Aubergine (shown in pink) and hAgo1 PAZ domain (shown in light blue) in the presence of 7mer RNA with 2'-O-methyl modification at the 3'-end.**

However, as shown in **Figure 63**, depicting Van der Waals surface representations of the 2'-O-methyl group on 3'-terminal nucleotide inserted into both Aubergine PAZ domain and the hAgo1 PAZ domain (Ago subfamily), there are steric clashes between the 2'-O-methyl group with the side chains of the residues lining the RNA binding pocket in hAgo1 PAZ domain, but no clashes in the binding pocket of the Aubergine PAZ domain.

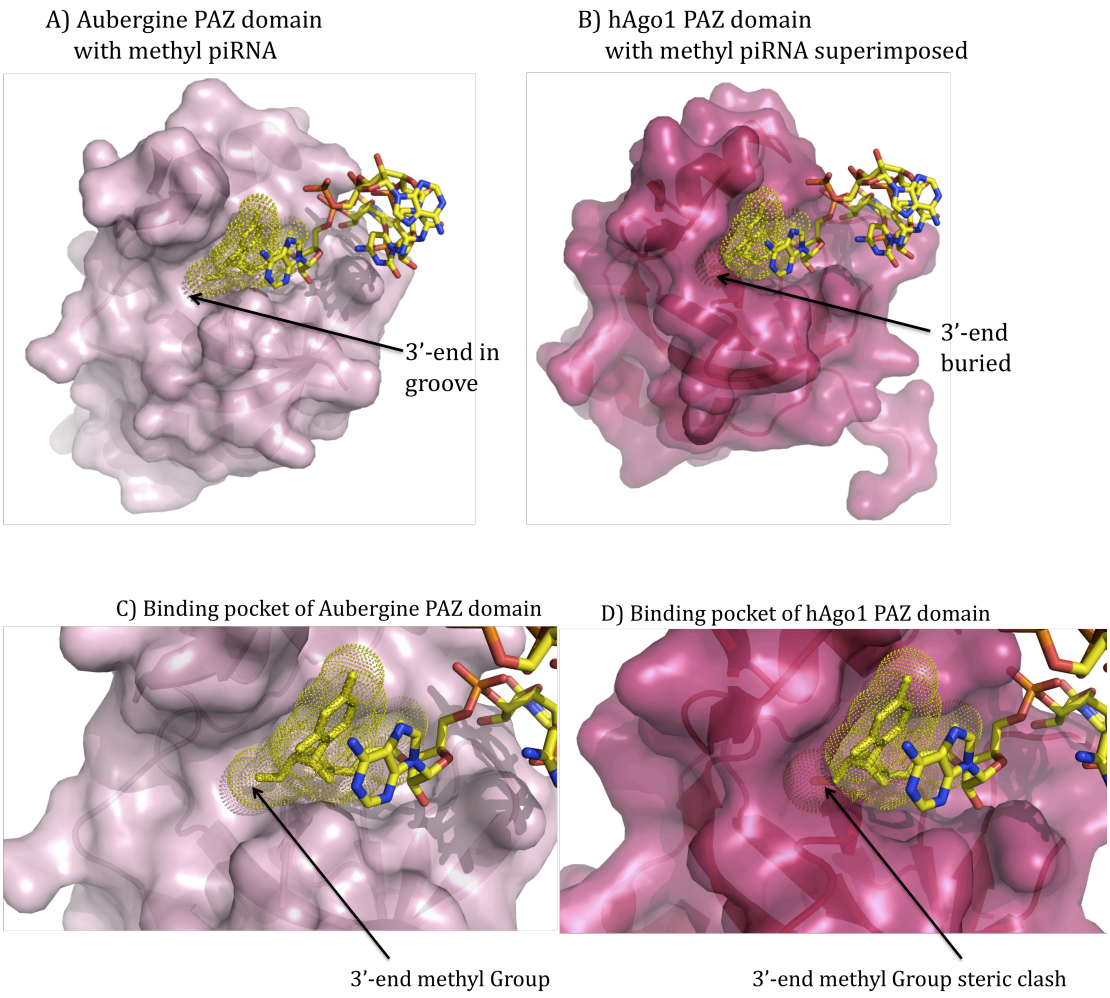



**Figure 63: Van der Waals surface representations of Aubergine and hAgo1 PAZ domain bound to 7mer RNA with 2'-O-methylation at 3'-end. A) Overall structure of Aubergine PAZ domain with methyl piRNA. B) Overall structure of hAgo1 with methyl piRNA superimposed. C) The binding pocket of Aubergine PAZ domain. D) The binding pocket of hAgo1 PAZ domain.**

The steric clashes observed in the hAgo1 RNA binding pocket contribute by the presence of several bulky amino acid residues including Valine and Phenyalanine on the αβ module and Tyrosine and Threonine on the β strand on the other side of the binding pocket; whereas in Aubergine PAZ domain there is a clear groove present to accommodate the 2'-O-methyl group at the 3'-end of the RNA. This could explain the observed subtle differences between the two structures.

Firstly, in the sequence alignment between the Piwi and Ago PAZ domain, the Aubergine PAZ domain has a Piwi subfamily specific ten amino acid insertion

between the β5 and β6 strands. As a result of missing this stretch of residues, the equivalent turn between the β5 and β6 strands in the hAgo1 PAZ domain is sharper and pushes the β6 strand further inward of the RNA binding pocket than the β6 strand in the Aubergine PAZ domain.

Secondly, the αβ module especially the β3-β4 hairpin of the Aubergine PAZ domain is shifted outwards of the RNA binding pocket compared to the β hairpin in the hAgo1 PAZ domain.

The result of the above two structure variations is that the RNA binding pocket of the Aubergine PAZ domain is wider than that of the hAgo1 PAZ domain and therefore accommodates and tolerates the addition of the 2'-O-methyl group.

### 3.4.4. Implication for other Argonautes

The 2'-O-methyl modification at the 3'-end is the defining feature of the mature piRNA. The function of the methylation is thought to be protecting it from uridylation and thus degradation. piRNAs only associate with the Piwi subfamily of Argonaute proteins and not the Ago subfamily of Argonaute proteins. The structure of the Aubergine PAZ domain in complex with the 3'- end of the piRNA provides insights into how the Piwi subfamily of Argonaute protein recognises its natural substrates. The widening of the RNA binding pocket and the presence of the hydrophobic binding cleft in the Aubergine PAZ domain are the main features that allow it to accommodate and preferentially bind to the 2'-O-methylated piRNAs at the 3' – end. The steric clashes observed for the insertion of the 2'-O-methyl group into the hAgo1 PAZ domain re-confirm that the Ago

subfamily PAZ domain strongly preferentially binds to the 3' – end of the non – methylated RNAs.

These structural features could perhaps be used to predict the potential to bind small RNAs with methyl modifications in other Argonaute proteins in other species. By superimposing the known PAZ domain structures with the *Drosophila* Aubergine PAZ/3'-end of piRNA complex, then substituting the *Drosophil*a Aubergine PAZ domain with the known PAZ domain structures in the presence of the 3'-end of the piRNA, gives the predicted compatibility model for PAZ domains with 2'- O – methylation at the 3'- end of the RNA. As shown in **Figure 64,** *Drosophila* Ago2 PAZ domain (pdb code – 1R6Z) (Song et al, 2003) has no steric clashes with the methyl group at the 3'- end of RNA, which implies that the natural substrates for *Drosophila* Ago2 PAZ domain could have 2'-O-methyl modification at the 3'- end despite the fact that *Drosophila* Ago2 belongs to the Ago subfamily of Argonaute proteins.

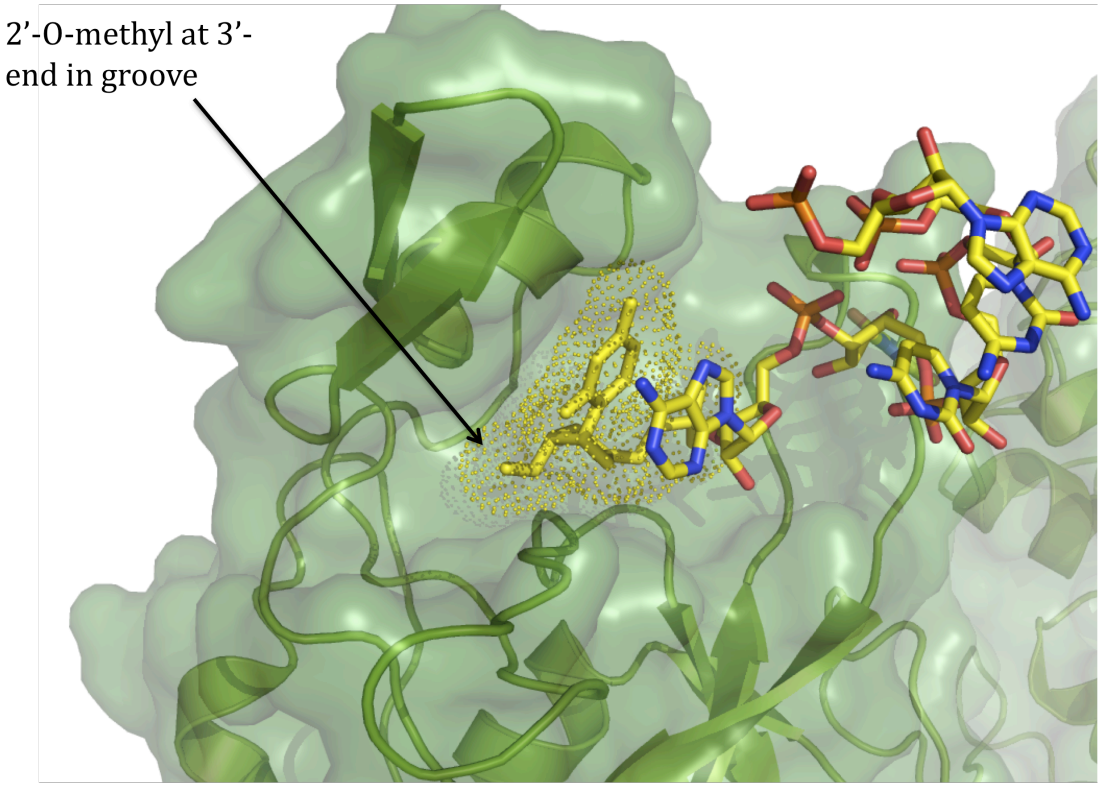



**Figure 64: Van der Waals radius representations of the dAgo2 PAZ domain (pdb code – 1R6Z) with 2'-O-methylated RNA at 3'-end superimposed in the RNA binding pocket.**

Previous studies show that siRNAs in Drosophila were in fact 2’-O-methylated at the 3’-end and they indeed bind to the *Drosophila* Ago2 (Horwich et al, 2007).

Despite having domain structural resemblance to the eukaryotic Argonaute proteins, little is known about the functions of prokaryotic Argonaute proteins and their natural DNA or RNA substrates. piRNAs so far have been only found in eukaryotes but by comparing the PAZ domain RNA binding sites, speculations can be drawn for potential prokaryotic Argonaute proteins and their substrates. As shown in **Figure 65** the *Pyrococcus furiosus* Argonaute PAZ domain (pdb code – 1U04) (Song et al, 2004) could potentially accommodate the presence of a 2’-O-methyl group at the 3’- end of the RNA without steric clashes.

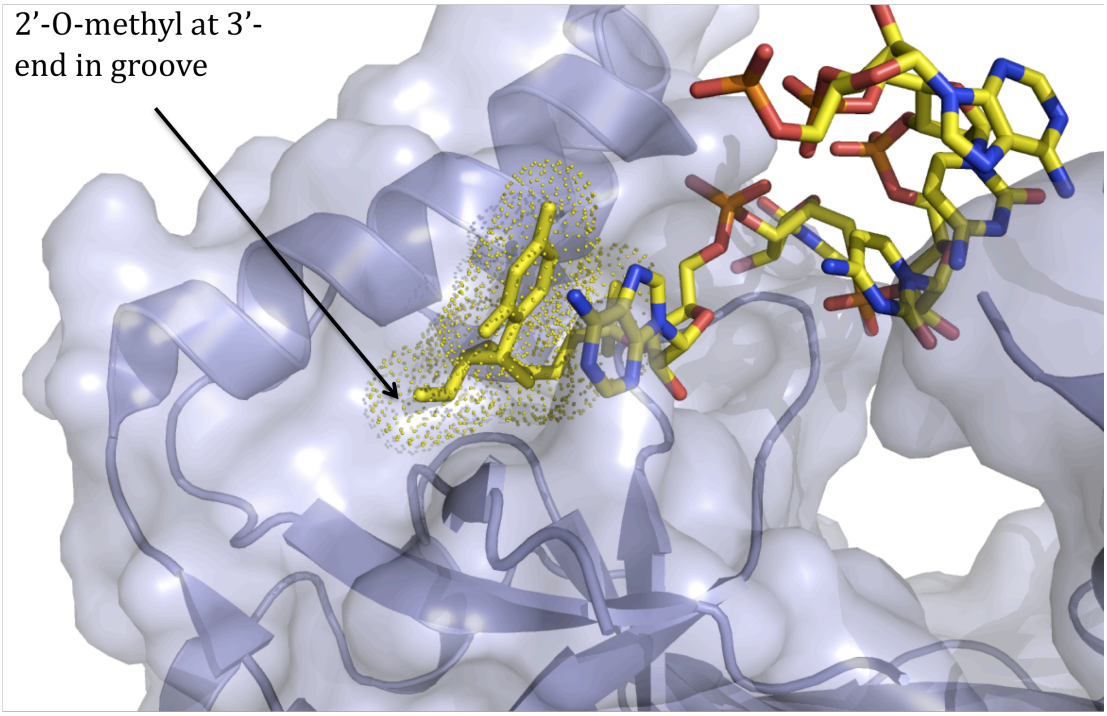



**Figure 65: Van Der Waals radius representations of the *Pyrococcus furiosus* Argonaute PAZ domain (pdb code - 1U04) with 2'-O-methylated RNA at 3'-end superimposed in the RNA binding pocket.**

This could mean that bacterial Argonaute proteins are implicated in binding and recognising RNA with 2'-O-methyl group at the 3'-end in certain pathways. As described in Section 3.1, in bacteria, the enzyme responsible for the 2'-O-methylation modification is thought to be involved in the RNA repair pathway and the Argonaute proteins might be involved during the repair process.

## 3.5. Biochemical Analysis of the PAZ domain from the Piwi Subfamily protein Aubergine

### 3.5.1. Introduction

To complement the structural studies of AubPAZ3$_{(279\text{-}398)}$, various biophysical techniques have been employed to characterise AubPAZ3$_{(279\text{-}398)}$ in solution, and to gain further understanding of the nature of substrate recognition for the PAZ domain of the Piwi subfamily of Argonaute protein.

### 3.5.2. Biophysical and biochemical studies of the Aubergine PAZ domain

#### 3.5.2.1. Analytical Gel Filtration

In the asymmetric unit of the crystal structure of AubPAZ3$_{(279-398)}$ bound to a 7mer RNA, it appeared to be a dimer in the crystal packing but with very small interacting interface. The oligomerisation state of the Aubergine PAZ domain could have implications for the piRNA pathway: as the identity of the nuclease responsible for trimming or slicing of the 3'- end of the piRNA to generate the mature piRNA remain elusive, enzymes with slicer activities such as Argonaute proteins could be implicated in this process. Thus further biochemical experiments were carried out to test the observation in the crystal structure and confirm the oligomerisation state of AubPAZ3$_{(279-398)}$ in solution.

The analytical superdex75 column was calibrated prior to running the AubPAZ3$_{(279-398)}$ sample. **Figure 66** shows the trace of the calibration and the experimental runs.

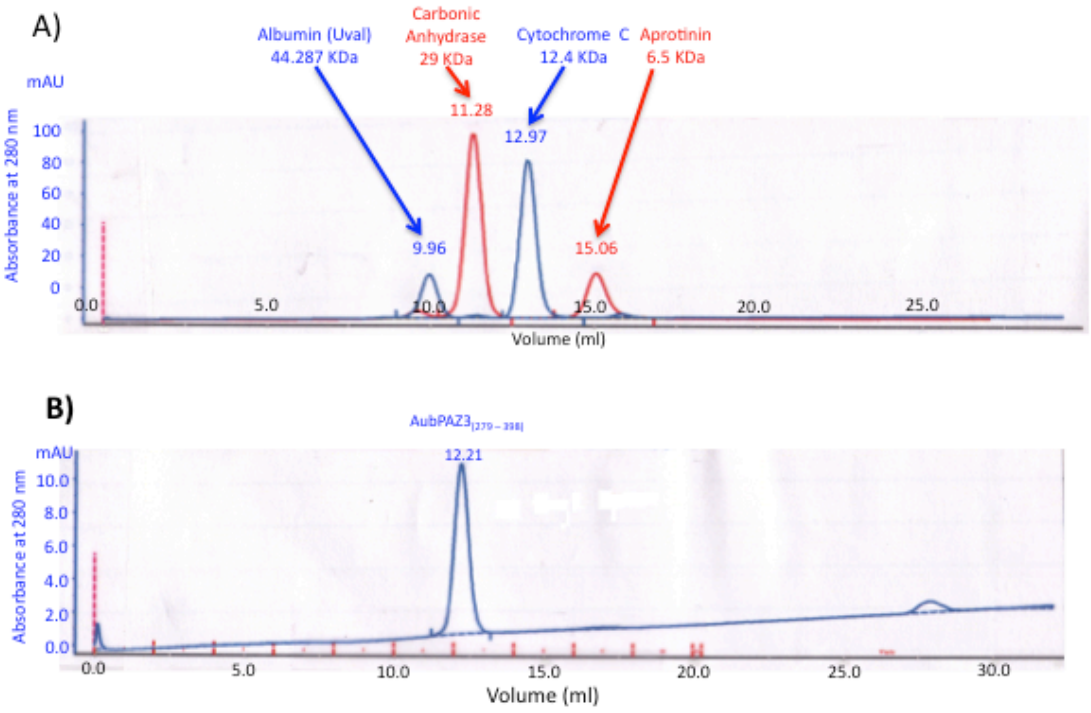



**Figure 66: A) Calibration run of the analytical gel filtration column superdex S75 using a mixture of protein standards. B) The elution profile of AubPAZ3$_{(279-398)}$ on the analytical gel filtration S75 column.**

**Table 13** and **Figure 67** show the calculation of the calibration

| Protein Marker | MW (Da) | $Log_{10}$ MW | Ve | (Ve - Vo) /(Vt - Vo) | |
|---|---|---|---|---|---|
| Albumin | 44,257 | 4.6459 | 9.96 | 0.1387 | Vo = 7.57ml |
| Carbonic Anhydrase | 29,000 | 4.4624 | 11.28 | 0.2196 | Vo = Void Volume |
| Cytochrome C | 12,400 | 4.0934 | 12.97 | 0.3233 | Vt = 24ml |
| Aprotinin | 6,500 | 3.8129 | 15.06 | 0.4515 | Vt = Column Volume |
| AubPAZ3 | | | 12.21 | 0.2767 | |

**Table 13: List of protein standards for gel filtration column calibration**

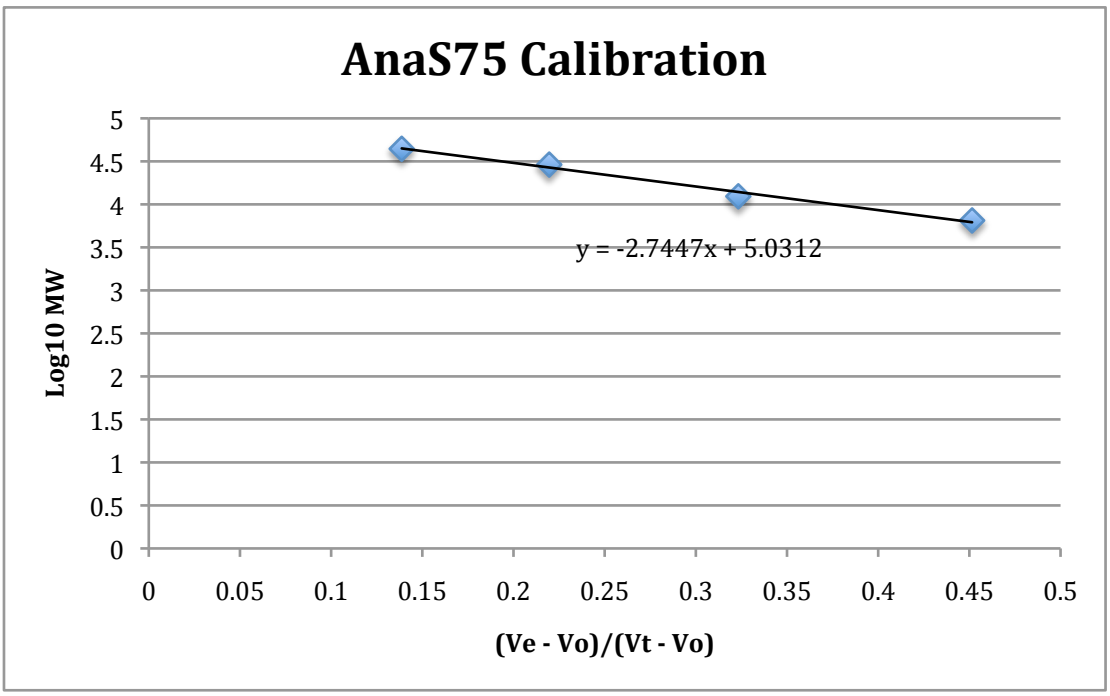



**Figure 67: The calculations of the molecular weight from the elution volume based on the calibration**

Therefore, line of best fit: y = -2.7447x + 5.0312

For AubPAZ3$_{(279\text{-}398)}$ :

$$\begin{aligned} \text{Log}_{10}\,\text{MW} &= -2.7447(0.2767) + 5.0312 \\ &= 4.2718 \end{aligned}$$

$$\begin{aligned} \text{MW} &= 10^{4.2718} \\ &= 18.698\ \text{KDa} \end{aligned}$$

The analytical gel filtration results showed that the AubPAZ3$_{(279\text{-}398)}$ protein run at an apparent larger molecular weight (18.7 kDa) than the theoretical molecular weight (14.2 kDa). The migration of protein in the gel filtration column depends on both the size, the shape of the protein as well as the interactions with the column. The apparent larger molecular weight could suggests either the presence of weak self association of AubPAZ3$_{(279\text{-}398)}$ protein, or non – globular shape of the protein thus faster migration.

### 3.5.2.2. Analytical Ultracentrifugation

To further examine and confirm the oligomerisation state of the AubPAZ3$_{(279\text{-}398)}$ in solution, analytical ultracentrifugation equilibrium was performed on 40 µM of sample (a stable concentration obtained from gel filtration) as described in Section 2.2.3.3. AUC determines the molecular weight independently of the shape of the molecule; this is in contrast to a gel filtration column in that the migration of the molecule is mass and shape dependant. **Figure 68** shows the absorbance (280 nm) versus rotor radius (cm) at 25,000 RPM and 30,000 RPM.

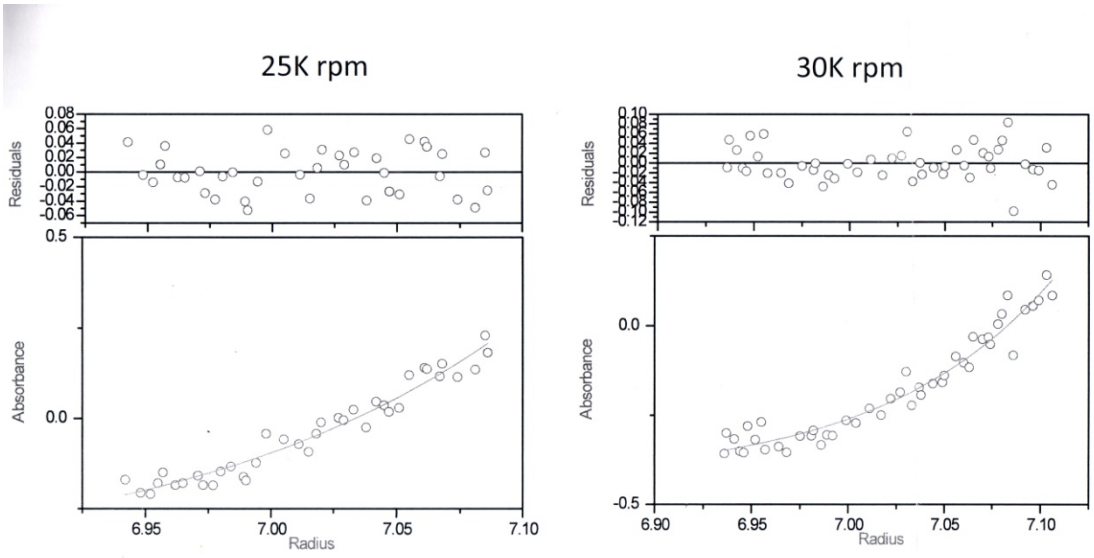



**Figure 68: AUC raw data. Left panel RPM = 25,000 and right panel RPM = 30,000**

The data was fitted to an ideal monodisperse model shown as a solid line. Residuals of the fit are plotted against rotor radius at the top of the graphs. A molecular weight of 14450 Da is consistent with monomeric particles in solution as the theoretical molecular weight was 14,280 Da. Therefore, it is unlikely that the dimer seen in the crystal also exists in solution or in the cell.

### 3.5.3. Interaction between Aubergine PAZ domain and methylated and unmethylated small RNA

In the crystal structure the Piwi subfamily of Argonaute PAZ domain specifically recognise the 3'-end of the piRNA with 2'-O-methylation while the Ago subfamily PAZ domain does not due to steric hindrance. However it also raises the question how does the Piwi subfamily of Argonaute PAZ domain recognise and associate with the piRNA with 2'-O-methylation modification over the non-methylated small RNAs? In contrast to piRNAs, the majority of the animal siRNAs (except for *Drosophila* siRNAs) and miRNAs do not have the 2'-O-methylation modification at the 3'-end. How does the Piwi subfamily of Argonaute protein distinguish between the two species? Thus several biophysical techniques were employed to investigate the interactions between the Piwi subfamily PAZ domain and the RNAs with or without 2'-O-methylation at the 3'- end to try to understand the nature of the substrate specificity for the Piwi subfamily of PAZ domain. Short RNAs (7 nucleotides) were chosen to avoid any self-annealing of RNA strands so that RNAs would remain in single-stranded form and mimic the natural piRNA PAZ domain substrates.

#### 3.5.3.1. By Isothermal Titration Calorimetry (ITC)

To understand the interaction between RNA and PAZ domain, isothermal titration calorimetry (ITC) experiments were performed between AubPAZ3$_{(279-398)}$ and 7mer RNA1 (5'-AACCUAC-Met3') as used in crystallisation. However the experiment was not successful as the heat measurements for protein RNA interaction was the same as the control experiment, which implies that either

there was no interactions between AubPAZ3$_{(279-398)}$ and RNA1; or the interaction itself does not generate heat change.

#### 3.5.3.2. By Band Shift Assay

Subsequently, EMSA (Electrophoretic Mobility Shift Assay) was employed to investigate the interaction between RNA (both with and without 2'-O-methylation at the 3'-end) and AubPAZ3$_{(279-398)}$. In the initial experiment the concentration range for AubPAZ3$_{(279-398)}$ was between 20 nM – 20 μM, and the RNA concentration in each reaction was 0.5 nM. **Figure 69** shows the results.

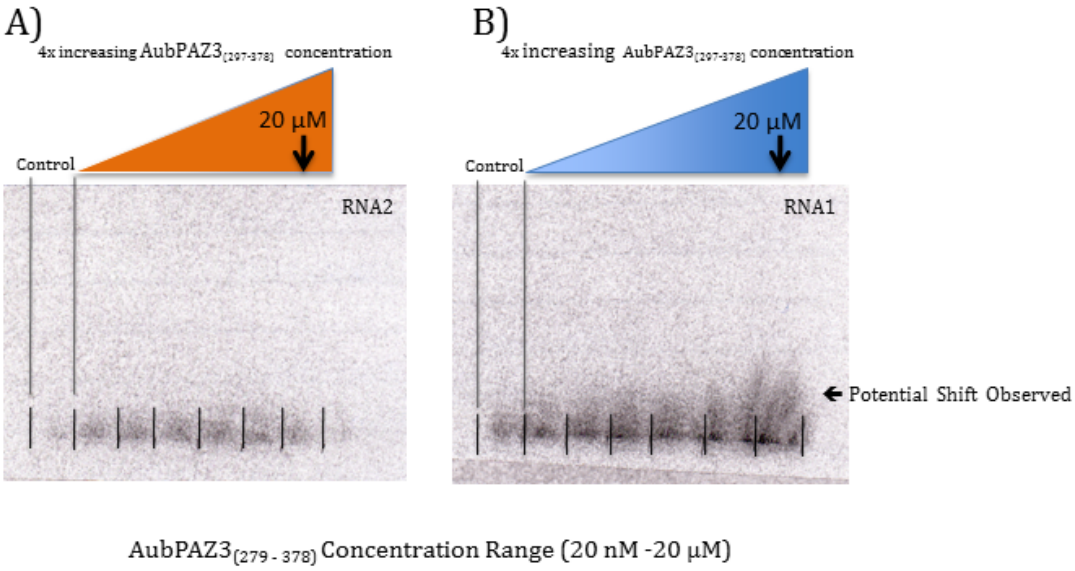



**Figure 69: A) EMSA result for AubPAZ3$_{(279-398)}$ with RNA2 (5'-AACCUAC-3') B) EMSA result for AubPAZ3$_{(279-398)}$ with RNA1 (5'-AACCUAC-Met3') that has 2'-O-methyl group at the 3'-end.**

Comparing the methylated form of RNA to the non – methylated form of RNA, there appeared to be a slight shift at the highest concentration (20 μM) of AubPAZ3$_{(279-398)}$ for RNA1 whereas no shifts were observed for RNA2, the non-methylated RNA.

The experiment was then repeated but with an even higher concentration range of AubPAZ3$_{(279-398)}$ . **Figure 70** shows the results.

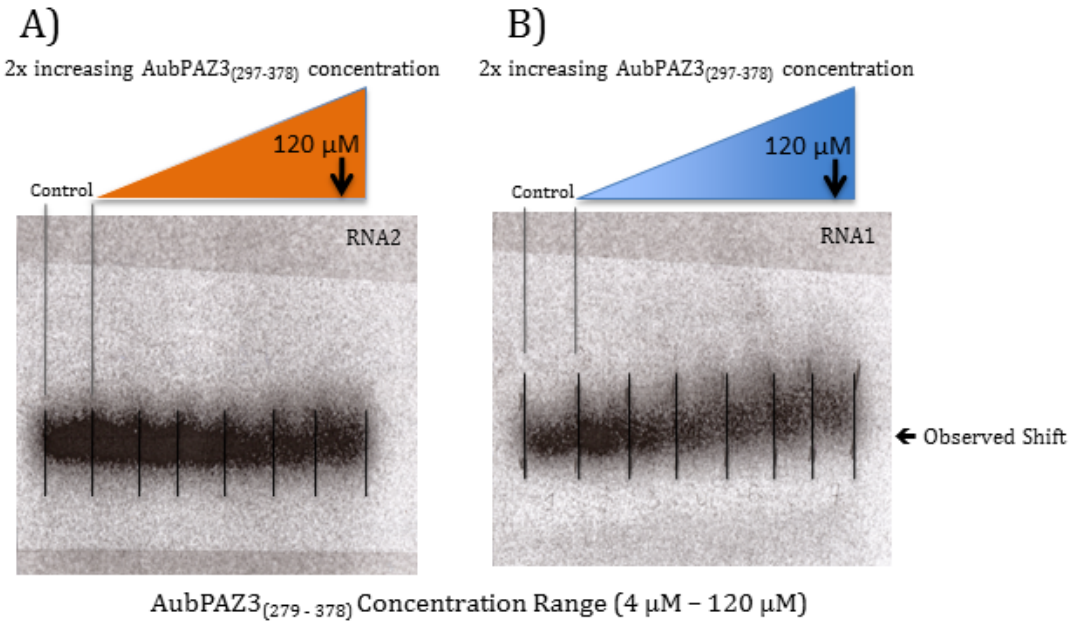



**Figure 70: A) EMSA result for AubPAZ3$_{(279-398)}$ at a higher concentration with RNA2 (5'-AACCUAC-3') B) EMSA result for AubPAZ3$_{(279-398)}$ with RNA1 (5'-AACCUAC-Met3')**

In this experiment the shifting of the band was observed more clearly than in the previous run. The result of this experiment was then used as the condition for crystallisation of the complex and subsequently for solving the structure. It also indicated that AubPAZ3$_{(279-398)}$ has a natural preference for the RNA with the 2'-O-methylation at the 3'-end which was the opposite of the hAgo1 PAZ domain, an Ago subfamily of PAZ domain that preferentially binds to RNA without the 2'-O-methyl group at the 3'- end. The small shift is probably due to the absence of the "caging" effect that generates larger shifts with longer nucleic acids.

#### 3.5.3.3. By Fluorescence Anisotropy

To further confirm the substrate preference of the AubPAZ3$_{(279-398)}$ protein, fluorescence anisotropy was employed to examine the binding of

AubPAZ3$_{(279-398)}$ to the 7-mer RNA with and without the 2'-O-methylation at the 3'-end (termed RNA3 (Fl-AACCUAC-Met) and RNA4 (Fl-AACCUAC) respectively) as described in Section 2.17.

As shown in **Figure 71**, AubPAZ3$_{(279-398)}$ binds to RNA3 (with estimated Kd of 20μM) more strongly than RNA4 (with estimated Kd of 50μM) and it confirmed the EMSA results that the Aubergine PAZ domain preferentially recognises and binds to the 3'-end of RNA with 2'-O-methyl modification.

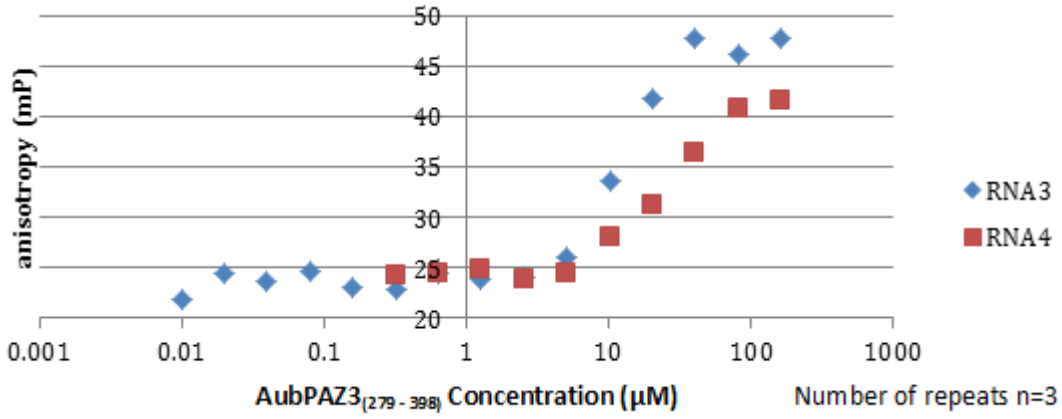


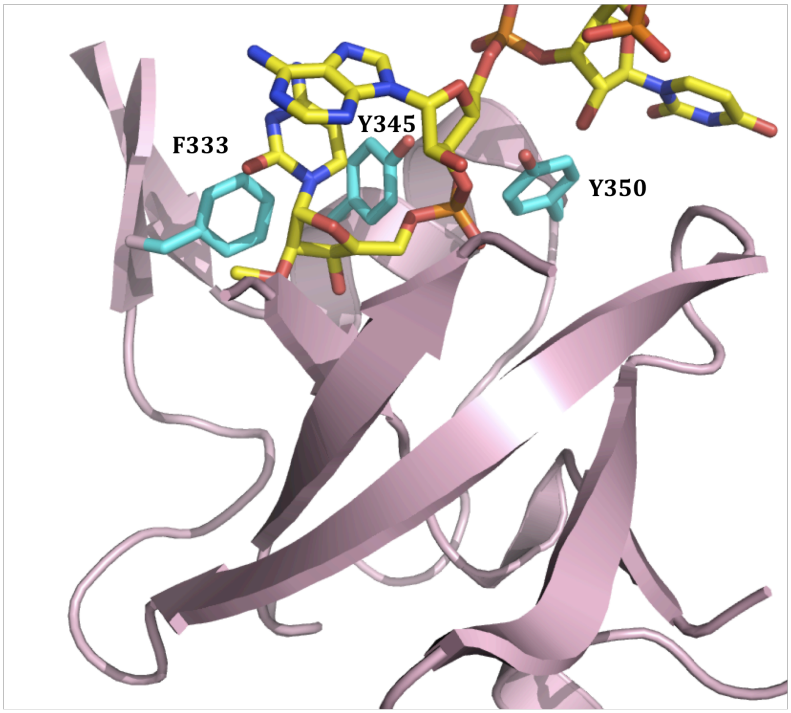

**Figure 71: Fluorescence anisotropy results for interactions between AubPAZ3$_{(279-398)}$ and RNA3 or RNA4 respectively.**

### 3.5.4. Production of AubPAZ3$_{(279-398)}$ mutants

Based on the structural information, contact mutants were designed to be used as tools to further study the properties of the PAZ domain.

### 3.5.4.1. Production of binding pocket mutants

Sequence alignment of homologous PAZ domain together with the crystal structure revealed several key conserved residues at the RNA binding pocket of AubPAZ3$_{(279\text{-}398)}$ as shown in **Figure 72**.




Aubergine PAZ domain RNA binding pocket

**Figure 72: The three key residues in the binding pocket of Aubergine PAZ domain involved in RNA substrate recognition.**

As described in Section 3.4, Phe333 (F333) and Tyr345 (Y345) were invariant residues across both the Ago and Piwi subfamilies of Argonaute protein and were involved in hydrophobic and hydrogen-bonding interactions with the 3'-end terminal nucleotide and the formation of the hydrophobic cleft in the RNA binding pocket. Tyr350 (Y350) was invariant across the Piwi subfamily of Argonaute proteins and highly conserved throughout both subfamilies of Argonaute proteins. It forms hydrogen bonds with both the phosphate backbone and the ribose group of the RNA.

For the purpose of examining the nature of the interaction between methylated RNA to the PAZ domain, three alanine mutants harbouring single, double and triple mutations (F333A, F333AY345A, F333AY345AY350A) respectively were cloned. This approach was also adopted for the dAgo2 PAZ domain structure studies (Song et al, 2003). However all three mutants were insoluble when expressed. This could be due to the nature of the mutation disrupting the protein fold too severely.

#### 3.5.4.2. Production of SumoPAZ

In order to improve the solubility of the PAZ mutants, the wild type AubPAZ3$_{(279-398)}$ was cloned into pOPINS vector (OPPF) and expressed as a fusion protein with His-SUMO tag denoted as SumoPAZ. SumoPAZ expression was induced by 0.1 mM IPTG at 18°C overnight in *E-coli* BL21 (DE3) strain. The purification of SumoPAZ scheme was the same as that of the wild type AubPAZ3$_{(279-398)}$ except that the His-Sumo tag was cleaved by Sumo Protease at room temperature for 3 hours. The yield for SumoPAZ was 5 mg per 1L prep, which is in order of 20 fold higher than for His-PAZ. **Figure 73** shows the trace of the final Gel filtration column and the SDS-PAGE gel of eluted fractions.

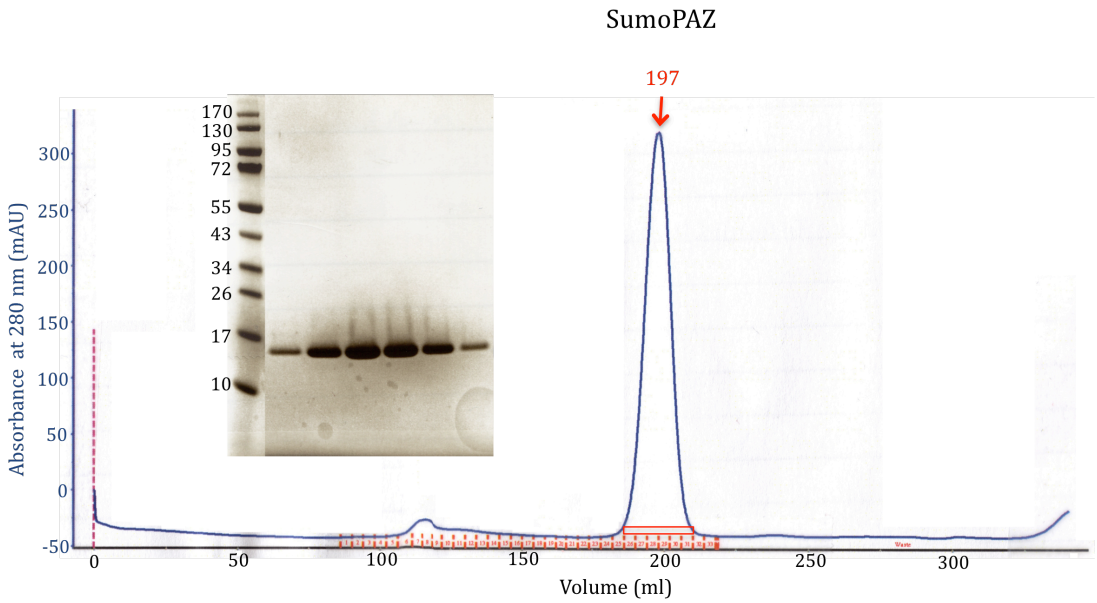



**Figure 73: Gel filtration trace and SDS-PAGE of SumoPAZ**

As shown in **Figure 73**, the SumoPAZ protein has a symmetric elution profile on the gel filtration column implying monodispersity of the sample. The elution volume (197 ml) is the same as the wild type AubPAZ3$_{(279-398)}$ gel filtration eluation volume and the purity of the sample confirmed by SDS-PAGE is comparable to that of wild type AubPAZ3$_{(279-398)}$ too.

### 3.5.4.3. Production of SumoPAZ mutants

As the wild type sumo AubPAZ3$_{(279-398)}$ protein appeared to be more soluble, it was assumed that the Sumo vector would increase the solubility of mutants. To further increase the chance of expressing soluble mutants, milder mutations were introduced to the key residues around the binding site. Based on the structure information with the biophysical studies, three classes of mutations were introduced with the following aims:

1) To disrupt the binding site.

Three Tyrosine to Phenylalanine mutants were generated to disrupt the binding of RNA to Aubergine PAZ domain. The Tyrosine to Phenylalanine mutation is less severe than Tyrosine to Alanine mutation as the only change from Tyrosine to Phenylalnine is the loss of hydroxyl group.

| Mutation | Location | Function |
|---|---|---|
| Tyr 313 (Y313F) | β1 strand | Forming hydrogen bond with the phosphate group on the terminal nucleotide |
| Tyr345 (Y345F) | α3 helix | Forming hydrogen bond with the Phosphate group on the terminal nucleotide |
| Tyr350 (Y350F) | α3 helix | Forming hydrogen bond with phosphate group on the terminal nucleotide and with the ribose group on the C5 nucleotide |

2) To artificially mimic the Ago binding pocket, to switch the binding preference from methylated to non-methylated RNA.

As Ago and Piwi Argonaute proteins have different preference for substrates as shown in Section 3.5.3.2 and 3.5.3.3, two mutants were made on the Aubergine PAZ domain to mimic the sequence in the Ago family PAZ domain to test if the preference of substrate could be switched.

| Mutations | Location | Function |
|---|---|---|
| Tyr 342 (Y342V) | α3 helix | The equivalent residue in Ago subfamily PAZ domain is mostly Valine |
| Tyr 342 (Y342I) | α3 helix | The equivalent residue in *Drosophila* Ago2 PAZ domain is Isoleucine |

3) Three mutants were made to investigate the role of the conserved loop that is specific to the Piwi subfamily of Argonaute proteins and stabilises the 'flipped' nucleotide. Asn314 (N314) is the key residue that stabilises the "flipped" nucleotide and its conformation was described in Section 3.4.2.5. Asn314 is never aspartic acid in any Piwi subfamily sequences and it is one of the most highly conserved residues, thus it gives rise to the questions of why it is so conserved and what role does it play in the structure?

| Mutation | Location | Function |
|---|---|---|
| Asn314 (N314C) | β1 strand | The equivalent of this residue in Ago PAZ domain is cysteine. |
| Asn314 (N314D) | β1 strand | To disrupt the amide bond to the "flipped" nucleotide |
| Asn314 (N314A) | β1 strand | To completely remove all contacts to the "flipped" nucleotide |

**Table 14** summarises the properties of the mutants.

| Mutants | Mutation | Cloned | Soluble as fusion protein | Stable after cleaving | Yield |
|---|---|---|---|---|---|
| SumoPAZMut1 | Y313F | Yes | Yes | No | - |
| SumoPAZMut2 | N314A | Yes | Yes | No | - |
| SumoPAZMut3 | N314C | Yes | Yes | No | - |
| SumoPAZMut4 | N314D | Yes | Yes | Yes | 8mg/ml |
| SumoPAZMut5 | Y342I | No | - | - | - |
| SumoPAZMut6 | Y342V | No | - | - | - |
| SumoPAZMut7 | Y345F | Yes | No | - | - |
| SumoPAZMut8 | Y350F | Yes | Yes | No | - |

**Table 14: Summary of the expression test for the SumoPAZ mutants**

**Table 14 and Figure 74** shows the solubility test results. The two mutants that mimic the binding pocket of the Ago subfamily of Argonaute PAZ domain were apparently toxic to *E coli* and could not be cloned. Unfortunately most of the rest of the mutants were either insoluble or unstable and precipitated after the cleavage of the Sumo tag. Only the SumoPAZMut4 (N314D) mutant was viable.

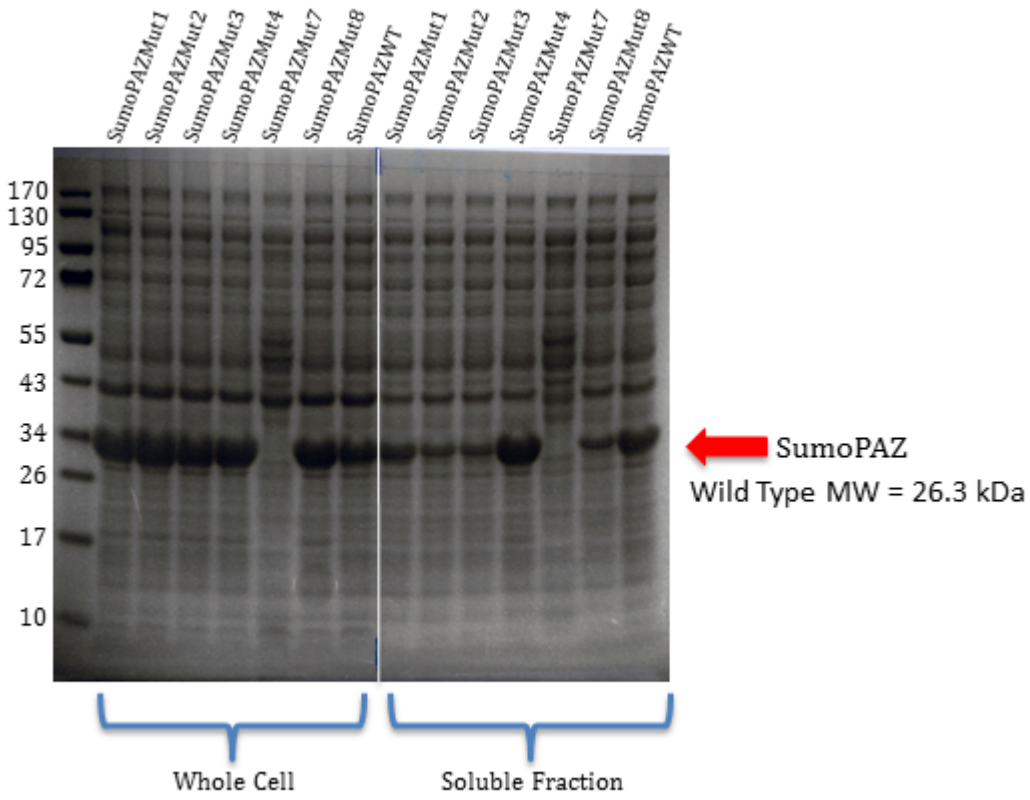



**Figure 74: expression test for the SumoPAZ mutants. The left hand panel shows the whole cell fractions and the right hand panel shows the soluble fractions**

### 3.5.5. Discussion

The biophysical studies showed that AubPAZ3$_{(279\text{-}398)}$ is a monomer in solution and it recognises and binds to the 3'-end of the RNA with 2'-O-methylation modification over the non-methylated RNA. The apparent preference for the methylated RNAs as substrates confirms with its biological function: as methylation on the 3'-end of the piRNA marks piRNA maturation and then associates and guides the Piwi subfamily of Argonaute protein to the target in the piRNA silencing pathway. To further examine the substrate specificity of AubPAZ3$_{(279\text{-}398)}$ and the structure features that were unique to the Piwi subfamily of the Argonaute PAZ domain, several attempts were made towards expressing and purifying multiple mutants of the AubPAZ3$_{(279\text{-}398)}$ protein.

Unfortunately the mutants were unstable and mostly insoluble even using the Sumo tag.

# 4. Discussion and Concluding Remarks

The field of RNA silencing has undergone astonishing development since its discovery in the early 1990s, not only owing to it being a relatively young field, but also for its important role in gene regulation across many species including human. Different classes of small RNAs and the components of the silencing effector complexes have been identified; subsequently various biogenesis and functional RNA silencing pathways have been gradually mapped out. They converge to a common theme for the silencing effector complex: a small RNA (20 nt – 30 nt in length) guides a member of the Argonaute protein family to drive the gene regulation.

piRNA is the most recently discovered class of small RNAs that function predominantly in germ cells for silencing mobile genetic elements such as transposons. Two of the key features that distinguish piRNAs from the other classes of small RNA are: 1) that they are derived from long single stranded RNA precursor and are ubiquitously methylated at the 3'-end on the 2'-OH group; 2) they associate exclusively with the Piwi subfamily of Argonaute proteins. Despite the developments in understanding piRNA pathway, not much is known about the piRNAs particularly 3'-end formation and regulation.

In this thesis, we hope to gain more understanding of the piRNA pathway through investigating two key aspects of the piRNA structurally and biochemically: firstly, the ubiquitous methylation at the 3'-end of the piRNA by enzyme Hen1 and secondly, piRNA specific recognition by the Piwi subfamily of Argonaute protein.

To address how the 3'-end of the piRNA is methylated by the methyltransferase Hen1, significant effort has been made towards producing functionally active, soluble and stable Hen1 constructs for structural studies. Three truncated constructs termed dHen1.4$_{(1-356)}$ , dHen1.12$_{(1-368)}$ and dHen1.14$_{(1-364)}$ were successfully expressed and purified to crystallization standard homogeneity. They also appeared to be stable and retained the enzymatic activities in *vitro*. The key findings revealed by methylation assays are that Hen1 protein is most active against 10 nt and 14 nt single stranded RNA substrates compared to the ~22 nt natural substrates; and the shortest soluble C-terminal truncated construct exhibits the highest activity against various lengths of small RNA substrates. This provides some interesting insights, as so far there have been no other reported *in vitro* methylation activity assays performed by a) non – fusion or truncated Hen1 protein; b) using RNA substrates that are shorter than 22 nt in length. Together with the solubility results of various dHen1 C-terminal truncation constructs, it is plausible to hypothesize that despite lacking known predicted domain in the C – terminal region of dHen1 protein, the C-terminal tail could be playing a structural role in making contacts with the N-terminal methyltransferase domain to keep it in a relatively inactive but also stable state. To achieve dHen1 activation fully, additional interacting partners may be involved to displace the C-terminal tail inhibition and protein could undergo significant conformation change.

Ultimately, solving the structure of dHen1 protein in complex with small RNA substrate would provide more mechanistic information on 2'-O-methylation at the 3'-end of the piRNA. In this thesis, it was shown that the 10 nt and 14 nt are the optimal lengthes for dHen1 *in vitro* activity thus suggesting the minimal

active complex for crystallization trials. This could serve as a starting point for future crystallization trails. Ultimately solving the structure of dHen1 bound to a small RNA substrate would provide insight into the piRNA 3'-end formation.

To investigate if Hen1 also functions in conjunction with other proteins, such as the Piwi subfamily of Argonaute proteins, in generating the 3' – end of the piRNA, full length Piwi subfamily of Argonaute proteins need to be produced. As they are known to be notoriously difficult to express (Rivas et al, 2005) (Miyoshi et al, 2005), an eukaryotic expression system such as a baculovirus/insect cells expression platform could be adopted instead of using the bacterial expression system. In addition, as described in section 1.2.4, Piwi subfamily of Argonaute proteins interact with the Tudor family proteins via the sDMA modification on Piwi Argonaute proteins. Thus co-expressing the Piwi subfamily of Argonautes and their corresponding interacting Tudor proteins could be another approach in producing soluble Piwi subfamily of Argonaute proteins.

In *vivo* studies could also be carried out to examine the effects on the level of siRNA methylation, piRNA methylation, the silencing of transposons and the cellular localization of the Hen1 protein, when Hen1 protein is truncated. Constructs could be designed based on our *in vitro* studies. Does the C-terminal tail regulate Hen1 *in vivo*? This would gain further insight into the dHen1 mechanism and complement the in *vitro* studies performed in this thesis.

piRNAs associate exclusively with the Piwi subfamily of Argonaute proteins and the PAZ domain of the Argonaute proteins recognises and binds to the 3'-end of the small RNAs. At the start of this study, there was no structural information on any part of any of the Piwi subfamily of Argonaute proteins. In this thesis, the

structure of Aubergine (a *Drosophila* Piwi subfamily of Argonaute protein) PAZ domain bound to the 3'- end of the piRNA with 2'-O-methylation provides the insight for specific RNA recognition by the Piwi subfamily of Argonaute protein.

The structure of Aubergine PAZ domain bound to a "7mer piRNA mimic" shows that the Piwi but not Ago subfamily of Argonaute protein can specifically recognize and bind to RNA with 2'-O-methylation at the 3'-end. Though the overall topology of Ago and Piwi subfamily of PAZ domain is mostly the same, the substrate specificity of Piwi subfamily PAZ domain arises from a) the presence of additional hydrophobic residues that create a more hydrophobic environment in the binding pocket to accommodate the additional methyl group; and b) the slight enlargement of the binding pocket compared to the Ago subfamily of PAZ domain that generates a hydrophobic groove for the 3'-end of piRNA to be anchored. The mode of recognition is completely non-sequence specific as all of the interactions observed between the RNA and the Aubergine PAZ domain are through interactions between the protein and the backbone of the RNA. This is in agreement with the fact that there are more than 1.5 million piRNA sequences. Our biochemical studies support the structural studies and demonstrate that the Aubergine PAZ domain preferentially binds to RNA with 2'-O-methylation at the 3'-end over RNAs without such modification. This confirms the substrate specificity of Piwi subfamily of Argonaute protein for piRNAs over other subclasses of small RNAs without the methylation modification at the 3'-end. Biochemical analysis can also be carried out between the piRNA substrates and the SumoPAZmut4 (N314D), a mutant that disrupts the interaction between the PAZ domain and the "flipped" base (3$^{rd}$ nucleotide from the 3'-end of the piRNA), to further gain insight into the significance of this structure feature

based on the structure of Aubergine PAZ domain bound to a "7mer piRNA mimic".

In Section 3.4, I show that the structures correlate to the known binding preferences for methylated and unmethylated RNAs. This could be therefore used to predict the possibility of methylation in less characterised systems such as bacteria. If this was to be taken further, molecular modeling using energy minimising tools might give more accurate predictions.

Earlier in 2011, two additional structures for the Piwi subfamily of PAZ domain bound to RNA have been solved. **Figure 75** shows the three Piwi subfamily PAZ domain structures. The first one is the crystal structure of human Hiwi1 bound to a 14 mer RNA duplex with 2 nt overhang at the 3'-end with 2'-O-methyl (PDB code 307V) (Tian et al, 2011). The second one is the NMR solution structure of mouse Miwi bound to 8 mer piRNA mimic (PDB code 2XFM) (Simon et al, 2011).

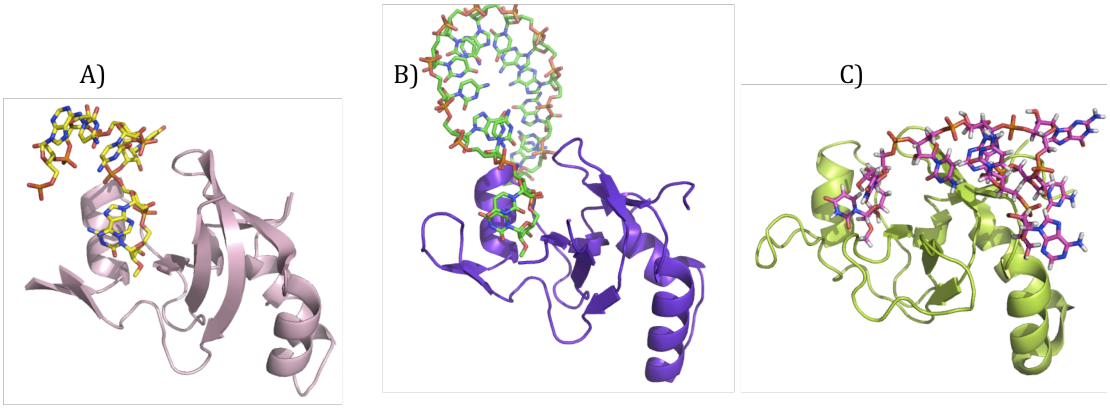
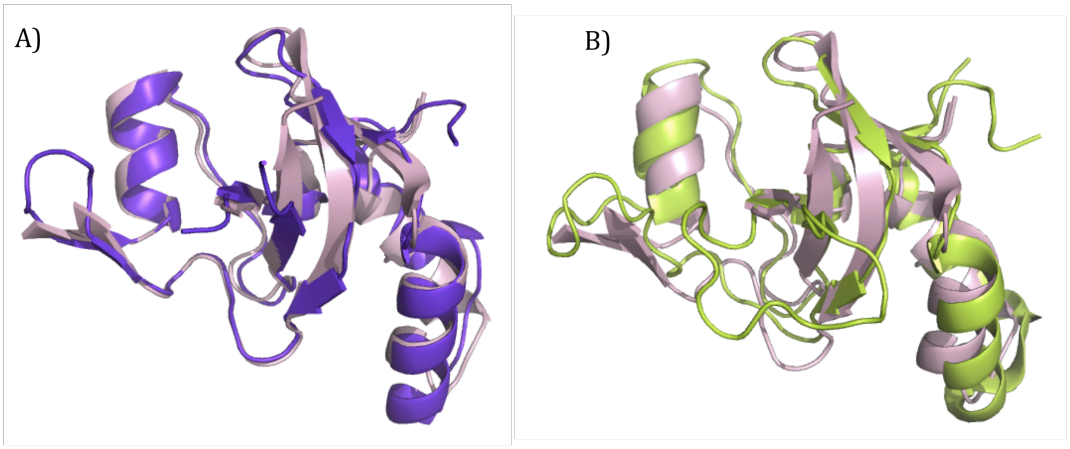



**Figure 75: Summary of the structures for Piwi subfamily PAZ domains. A) The crystal structure of the Aubergine PAZ domain bound to a single stranded 7 mer piRNA mimic in (this thesis). B) The crystal structure of Hiwi1 PAZ domain bound to a 14 mer RNA duplex with 2 nt overhang at the 3'-end (only one half of the RNA duplex is shown). C) The solution NMR structure of the Miwi PAZ domain bound to a single stranded 8 mer piRNA mimic.**

These two studies also compared the binding of methylated and unmethylated RNA to the Piwi PAZ domains. The results are in agreement with the biochemical studies conducted in this study in section 3.5.3 – that RNA with 2'-O-methylation at the 3'-end binds to the Piwi subfamily of PAZ domain slightly more tightly than the non-methylated RNA.

Superimposed structure of Aubergine, Hiwi1 and Miwi shown in **Figure 76** demonstrates that the three proteins have almost identical confirmation, suggesting that they share the same mechanisms in recognising and binding to small RNA substrates. The Miwi PAZ domain superimposes slightly less well.




Aubergine and Hiwi1 PAZ domain superimposed

Aubergine and Miwi PAZ domain superimposed

**Figure 76: Superimposed PAZ domains. A) Aubergine (pink) and Hiwi1 (purple) PAZ domain superimposed. B) Aubergine (pink) and Miwi (yellow) PAZ domain superimposed.**

The main difference between the three structures of Piwi subfamily PAZ domains / RNA complexes is the trajectory of the RNA as shown in **Figure 77**.

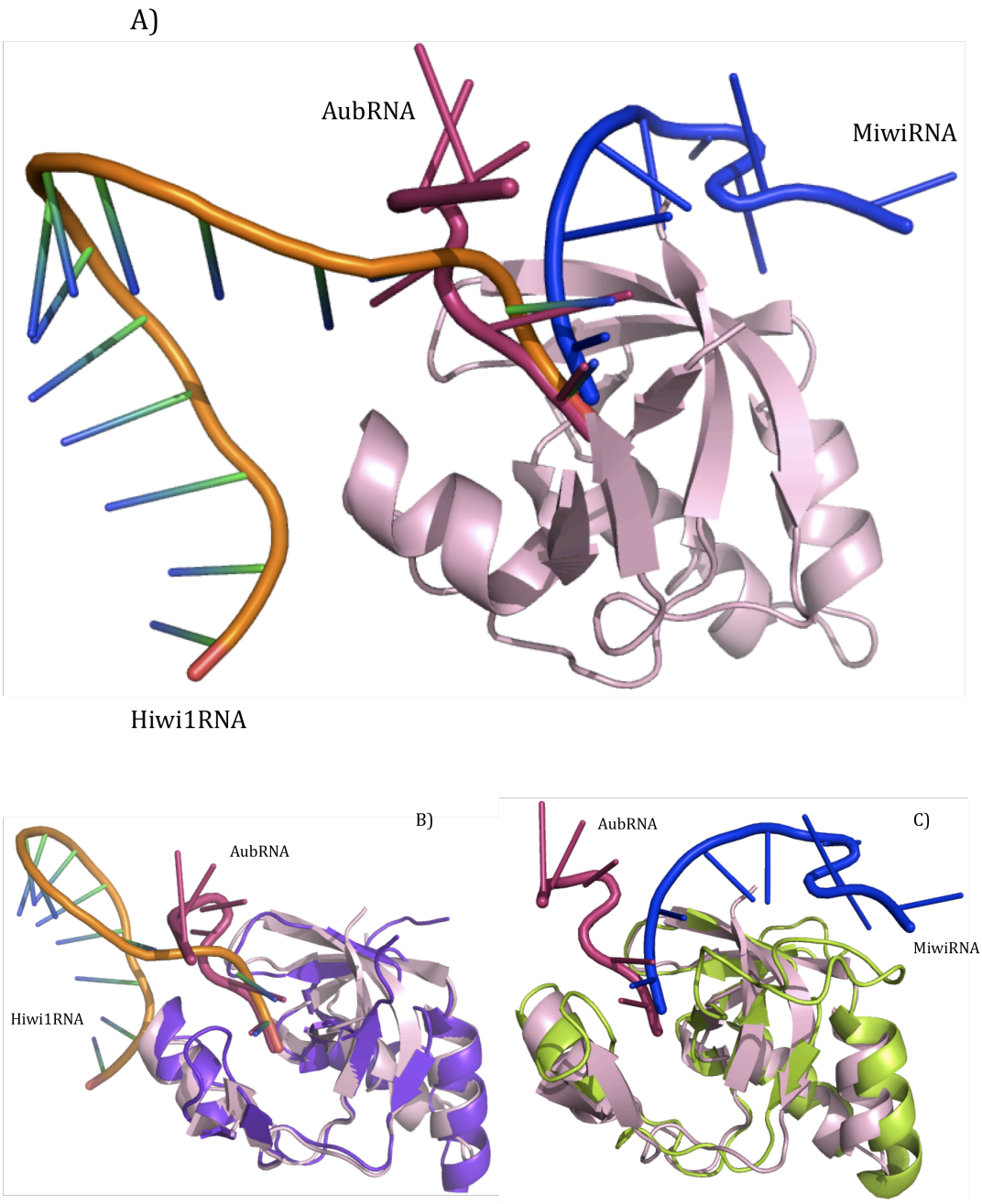



**Figure 77: RNA trajectories in the Piwi subfamily of PAZ domains. A) Three PAZ/RNA complex structures superimposed. Only the Aubergine PAZ domain is shown. B) The RNA binding pockets of Aubergine (pink) and Hiwi1 (purple) PAZ domain superimposed. C) The RNA binding pockets of Aubergine (pink) and Miwi (yellow) PAZ domain superimposed.**

Aubergine PAZ domain and Hiwi1 PAZ domain share almost identical protein structure and as shown in **Figure 77B** the first two nucleotides at the 3'-end of the RNA also superimposed well, confirming that the mechanism for the recognition of the RNA substrate is in agreement. However, from the third

nucleotide counting from the 3'-end of the RNA, the two RNA molecules show significant difference in their trajectories. The involvement of conserved residues and the specific recognition of the last two nucleotides at the 3'-end of RNA are very similar in both structures. Tyr318, Tyr345, Tyr313, Tyr 350 and Tyr346 in Aubergine (Tyr317, Tyr345, Tyr312, Tyr350 and Tyr346 in Hiwi1) are involved in forming hydrogen bonds with the last two nucleotides at the 3'-end of the RNA and these interactions are in agreement between the two structures. However the specific recognition of the "flipped" base nucleotide (the third nucleotide from the 3'-end) observed in AubPAZ – RNA complex structure is not observed in the Hiwi1PAZ – RNA structure. The 3$^{rd}$ nucleotide from the 3'-end does not superimpose well between the two structures. The residues involved in stabilising the "flipped" base nucleotide in Aubergine are Asn314 and Arg349, both are conserved across the Piwi subfamily of PAZ domain, but the equivalent residues in Hiwi1PAZ domain (Asn313 and Gln350) are not involved in forming any interactions with the RNA bound to the Hiwi1PAZ domain. Thus the discrepancy in the RNA trajectory beyond the 2$^{nd}$ nucleotide from the 3'-end could be explained by the RNA molecule bound to the Hiwi1PAZ domain was not a single stranded RNA as in the case of piRNA, but a RNA duplex with 2 nt overhang at the 3'-end; whereas in Aubergine PAZ domain, the 7 mer RNA bound is carefully chosen to remain single stranded and mimic the most authentic piRNA. [The Hiwi1PAZ-RNA complex is not authentic because the RNA is released from the PAZ domain when double-stranded (bound to a target), therefore only single-stranded bound is the biological relevant structure.]

The Miwi PAZ domain as well as the bound RNA show further differences in the structure, shown in **Figure 77C**. The MiwiPAZ–RNA structure is less well defined

in the RNA binding site, due to broadening of the NMR resonances. In addition, there are no intermolecular NMR restraints involving the second nucleotide from the 3'-end, making it difficult to place. The specific interactions between the PAZ domain and the last two nucleotides at the 3'-end of the RNA in both AubPAZ and Hiwi1PAZ structure are not observed in the MiwiPAZ-RNA structure. The conversed residues in Aubergine PAZ domain that are involved in 3'-end RNA binding include Tyr318, Tyr345, Tyr313, Tyr 350 and Tyr346. The equivalent conserved residues in MiwiPAZ (Tyr318, Tyr345, Tyr313, Tyr 350 and Tyr346 in Miwi) are *not* involved in recognition of the 3'-end of the RNA. In fact, Tyr350 appears to clash with a non-bridging oxygen on the sugar-phosphate backbone.

In summary, the three Piwi subfamily of PAZ domains show the same overall structure and the mode of substrate recognition. The AubPAZ and Hiwi1PAZ domain show almost identical protein structures, and the recognition of the two 3' end nucleotides is very similar. The MiwiPAZ domain structure, although an authentic single-stranded RNA structure, displays more differences which may reflect fewer restraints involving the RNA.

# 5. References


Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58**(Pt 11)**:** 1948-1954

Ameres SL, Horwich MD, Hung JH, Xu J, Ghildiyal M, Weng Z, Zamore PD (2010) Target RNA-directed trimming and tailing of small silencing RNAs. *Science* **328**(5985)**:** 1534-1539

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo JJ, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**(7099)**:** 203-207

Aravin AA, Hannon GJ, Brennecke J (2007a) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**(5851)**:** 761-764

Aravin AA, Klenov MS, Vagin VV, Bantignies F, Cavalli G, Gvozdev VA (2004) Dissection of a natural RNA silencing process in the Drosophila melanogaster germ line. *Mol Cell Biol* **24**(15)**:** 6742-6750

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T (2003) The small RNA profile during Drosophila melanogaster development. *Dev Cell* **5**(2)**:** 337-350

Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol* **11**(13)**:** 1017-1027

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**(6)**:** 785-799

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007b) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**(5825)**:** 744-747

Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev* **22**(20)**:** 2773-2785

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, Conte D, Jr., Luo S, Schroth GP, Carrington JC, Bartel DP, Mello CC (2008) PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol Cell* **31**(1)**:** 67-78

Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* **67**(Pt 4)**:** 271-281

Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**(6818)**:** 363-366

Biemont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. *Nature* **443**(7111)**:** 521-524

Bohmert K, Camus I, Bellini C, Bouchez D, Caboche M, Benning C (1998) AGO1 defines a novel locus of Arabidopsis controlling leaf development. *EMBO J* **17**(1)**:** 170-180

Boland A, Tritschler F, Heimstadt S, Izaurralde E, Weichenrieder O (2010) Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein. *EMBO Rep* **11**(7)**:** 522-527

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* **123**(7)**:** 1279-1291

Boswell RE, Mahowald AP (1985) tudor, a gene required for assembly of the germ plasm in Drosophila melanogaster. *Cell* **43**(1)**:** 97-104

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**(6)**:** 1089-1103

Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**(12)**:** 1957-1966

Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**(4)**:** 503-514

Carmell MA, Xuan Z, Zhang MQ, Hannon GJ (2002) The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev* **16**(21)**:** 2733-2742

Chan CM, Zhou C, Huang RH (2009) Reconstituting bacterial RNA repair and modification in vitro. *Science* **326**(5950)**:** 247

Chapman EJ, Carrington JC (2007) Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* **8**(11)**:** 884-896

Chen C, Jin J, James DA, Adams-Cioaba MA, Park JG, Guo Y, Tenaglia E, Xu C, Gish G, Min J, Pawson T (2009) Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. *Proc Natl Acad Sci U S A* **106**(48)**:** 20336-20341

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**(7051)**:** 740-744

Chuma S, Hosokawa M, Kitamura K, Kasai S, Fujioka M, Hiyoshi M, Takamune K, Noce T, Nakatsuji N (2006) Tdrd1/Mtr-1, a tudor-related gene, is essential for male germ-cell differentiation and nuage/germinal granule formation in mice. *Proc Natl Acad Sci U S A* **103**(43)**:** 15894-15899

Chung WJ, Okamura K, Martin R, Lai EC (2008) Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Curr Biol* **18**(11)**:** 795-802

Cohen SX, Ben Jelloul M, Long F, Vagin A, Knipscheer P, Lebbink J, Sixma TK, Lamzin VS, Murshudov GN, Perrakis A (2008) ARP/wARP and molecular replacement: the next generation. *Acta Crystallogr D Biol Crystallogr* **64**(Pt 1)**:** 49-60

Cowtan K (2000) General quadratic functions in real and reciprocal space and their application to likelihood phasing. *Acta Crystallogr D Biol Crystallogr* **56**(Pt 12)**:** 1612-1621

Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62**(Pt 9)**:** 1002-1011

Cox DN, Chao A, Baker J, Chang L, Qiao D, Lin H (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* **12**(23)**:** 3715-3727

Cox DN, Chao A, Lin H (2000) piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127**(3)**:** 503-514

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, Hannon GJ, Brennecke J (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature* **453**(7196)**:** 798-802

Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R, Matthews N, Berezikov E, Ketting RF, Tavare S, Miska EA (2008) Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. *Mol Cell* **31**(1)**:** 79-90

Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**(7014)**:** 231-235

Desset S, Buchon N, Meignin C, Coiffet M, Vaury C (2008) In Drosophila melanogaster the COM locus directs the somatic silencing of two retrotransposons through both Piwi-dependent and -independent pathways. *PLoS One* **3**(2)**:** e1526

Elbashir SM, Lendeckel W, Tuschl T (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**(2)**:** 188-200

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**(Pt 12 Pt 1)**:** 2126-2132

Eulalio A, Huntzinger E, Izaurralde E (2008) Getting to the root of miRNA-mediated gene silencing. *Cell* **132**(1)**:** 9-14

Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9**(2)**:** 102-114

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**(6669)**:** 806-811

Forstemann K, Tomari Y, Du T, Vagin VV, Denli AM, Bratu DP, Klattenhoff C, Theurkauf WE, Zamore PD (2005) Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* **3**(7)**:** e236

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z, Zamore PD (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* **320**(5879)**:** 1077-1081

Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**(2)**:** 94-108

Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**(7099)**:** 199-202

Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. *Cell Mol Life Sci* **66**(23)**:** 3727-3742

Goodier JL, Kazazian HH, Jr. (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**(1)**:** 23-35

Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R (2005) Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* **123**(4)**:** 631-640

Grivna ST, Beyret E, Wang Z, Lin H (2006a) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**(13)**:** 1709-1714

Grivna ST, Pyhtila B, Lin H (2006b) MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc Natl Acad Sci U S A* **103**(36)**:** 13415-13420

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* **315**(5818)**:** 1587-1590

Haase AD, Fenoglio S, Muerdter F, Guzzardo PM, Czech B, Pappin DJ, Chen C, Gordon A, Hannon GJ (2010) Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. *Genes Dev* **24**(22)**:** 2499-2504

Hamilton AJ, Baulcombe DC (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**(5441)**:** 950-952

Hammond SM, Bernstein E, Beach D, Hannon GJ (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404**(6775)**:** 293-296

Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18**(24)**:** 3016-3027

Hannon GJ (2002) RNA interference. *Nature* **418**(6894)**:** 244-251

Hartig JV, Esslinger S, Bottcher R, Saito K, Forstemann K (2009) Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. *EMBO J* **28**(19)**:** 2932-2944

Henz SR, Cumbie JS, Kasschau KD, Lohmann JU, Carrington JC, Weigel D, Schmid M (2007) Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol* **144**(3)**:** 1247-1255

Herr AJ, Baulcombe DC (2004) RNA silencing pathways in plants. *Cold Spring Harb Symp Quant Biol* **69:** 363-370

Horwich MD, Li C, Matranga C, Vagin V, Farley G, Wang P, Zamore PD (2007) The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* **17**(14)**:** 1265-1272

Hosokawa M, Shoji M, Kitamura K, Tanaka T, Noce T, Chuma S, Nakatsuji N (2007) Tudor-related proteins TDRD1/MTR-1, TDRD6 and TDRD7/TRAP: domain composition, intracellular localization, and function in male germ cells in mice. *Dev Biol* **301**(1)**:** 38-52

Houwing S, Berezikov E, Ketting RF (2008) Zili is required for germ cell differentiation and meiosis in zebrafish. *EMBO J* **27**(20)**:** 2702-2711

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, Plasterk RH, Hannon GJ, Draper BW, Ketting RF (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**(1)**:** 69-82

Huang Y, Ji L, Huang Q, Vassylyev DG, Chen X, Ma JB (2009) Structural insights into mechanisms of the small RNA methyltransferase HEN1. *Nature* **461**(7265)**:** 823-827

Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**(5531)**:** 834-838

Jain R, Shuman S (2009) Bacterial Hen1 is a 3' terminal RNA ribose 2'-O-methyltransferase component of a bacterial RNA repair cassette. *RNA* **16**(2)**:** 316-323

Jain R, Shuman S (2011) Active site mapping and substrate specificity of bacterial Hen1, a manganese-dependent 3' terminal RNA ribose 2'O-methyltransferase. *RNA* **17**(3)**:** 429-438

Kamminga LM, Luteijn MJ, den Broeder MJ, Redl S, Kaaij LJ, Roovers EF, Ladurner P, Berezikov E, Ketting RF (2010) Hen1 is required for oocyte development and piRNA stability in zebrafish. *EMBO J* **29**(21)**:** 3688-3700

Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* **100**(11)**:** 6569-6574

Katiyar-Agarwal S, Morgan R, Dahlbeck D, Borsani O, Villegas A, Jr., Zhu JK, Staskawicz BJ, Jin H (2006) A pathogen-inducible endogenous siRNA in plant immunity. *Proc Natl Acad Sci U S A* **103**(47)**:** 18002-18007

Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, Okada TN, Siomi MC, Siomi H (2008) Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453**(7196)**:** 793-797

Kawaoka S, Hayashi N, Suzuki Y, Abe H, Sugano S, Tomari Y, Shimada T, Katsuma S (2009) The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA* **15**(7)**:** 1258-1264

Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. *Science* **303**(5664)**:** 1626-1632

Kim DH, Saetrom P, Snove O, Jr., Rossi JJ (2008) MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc Natl Acad Sci U S A* **105**(42)**:** 16230-16235

Kim VN (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6**(5)**:** 376-385

Kirino Y, Kim N, de Planell-Saguer M, Khandros E, Chiorean S, Klein PS, Rigoutsos I, Jongens TA, Mourelatos Z (2009a) Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nat Cell Biol* **11**(5)**:** 652-658

Kirino Y, Mourelatos Z (2007a) Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. *Nat Struct Mol Biol* **14**(4)**:** 347-348

Kirino Y, Mourelatos Z (2007b) The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. *RNA* **13**(9)**:** 1397-1401

Kirino Y, Vourekas A, Kim N, de Lima Alves F, Rappsilber J, Klein PS, Jongens TA, Mourelatos Z (2010) Arginine methylation of vasa protein is conserved across phyla. *J Biol Chem* **285**(11)**:** 8148-8154

Kirino Y, Vourekas A, Sayed N, de Lima Alves F, Thomson T, Lasko P, Rappsilber J, Jongens TA, Mourelatos Z (2009b) Arginine methylation of Aubergine mediates Tudor binding and germ plasm localization. *RNA* **16**(1)**:** 70-78

Klattenhoff C, Bratu DP, McGinnis-Schultz N, Koppetsch BS, Cook HA, Theurkauf WE (2007) Drosophila rasiRNA pathway mutations disrupt embryonic axis specification through activation of an ATR/Chk2 DNA damage response. *Dev Cell* **12**(1)**:** 45-55

Kojima K, Kuramochi-Miyagawa S, Chuma S, Tanaka T, Nakatsuji N, Kimura T, Nakano T (2009) Associations between PIWI proteins and TDRD1/MTR-1 are critical for integrated subcellular localization in murine male germ cells. *Genes Cells* **14**(10)**:** 1155-1165

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, Hata K, Li E, Matsuda Y, Kimura T, Okabe M, Sakaki Y, Sasaki H, Nakano T (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22**(7)**:** 908-917

Kurth HM, Mochizuki K (2009) 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. *RNA* **15**(4)**:** 675-685

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science* **294**(5543)**:** 853-858

Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of Drosophila microRNA genes. *Genome Biol* **4**(7)**:** R42

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C,

Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822)**:** 860-921

Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**(5543)**:** 858-862

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE (2006) Characterization of the piRNA complex from rat testes. *Science* **313**(5785)**:** 363-367

Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**(5543)**:** 862-864

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**(5)**:** 843-854

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**(6956)**:** 415-419

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004a) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23**(20)**:** 4051-4060

Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW (2004b) Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117**(1)**:** 69-81

Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, Kittler EL, Zapp ML, Klattenhoff C, Schulz N, Theurkauf WE, Weng Z, Zamore PD (2009) Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**(3)**:** 509-521

Li J, Yang Z, Yu B, Liu J, Chen X (2005) Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. *Curr Biol* **15**(16)**:** 1501-1507

Li MA, Alls JD, Avancini RM, Koo K, Godt D (2003) The large Maf factor Traffic Jam controls gonad morphogenesis in Drosophila. *Nat Cell Biol* **5**(11)**:** 994-1000

Lin H, Spradling AC (1997) A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development* **124**(12)**:** 2463-2476

Lingel A, Simon B, Izaurralde E, Sattler M (2003) Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. *Nature* **426**(6965)**:** 465-469

Liu H, Wang JY, Huang Y, Li Z, Gong W, Lehmann R, Xu RM (2010a) Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev* **24**(17)**:** 1876-1881

Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**(5689)**:** 1437-1441

Liu K, Chen C, Guo Y, Lam R, Bian C, Xu C, Zhao DY, Jin J, MacKenzie F, Pawson T, Min J (2010b) Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *Proc Natl Acad Sci U S A* **107**(43)**:** 18398-18403

Liu Q, Rand TA, Kalidas S, Du F, Kim HE, Smith DP, Wang X (2003) R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* **301**(5641)**:** 1921-1925

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) Elucidation of the small RNA component of the transcriptome. *Science* **309**(5740)**:** 1567-1569

Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. *Science* **303**(5654)**:** 95-98

Ma JB, Ye K, Patel DJ (2004) Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**(6989)**:** 318-322

Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ (2005) Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature* **434**(7033)**:** 666-670

MacRae IJ, Ma E, Zhou M, Robinson CV, Doudna JA (2008) In vitro reconstitution of the human RISC-loading complex. *Proc Natl Acad Sci U S A* **105**(2)**:** 512-517

Macrae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA (2006) Structural basis for double-stranded RNA processing by Dicer. *Science* **311**(5758)**:** 195-198

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137**(3)**:** 522-535

Maniatis T (2009) Molecular cloning: A Laboratory Manual.

Marques JT, Kim K, Wu PH, Alleyne TM, Jafari N, Carthew RW (2009) Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. *Nat Struct Mol Biol* **17**(1)**:** 24-30

Matranga C, Tomari Y, Shin C, Bartel DP, Zamore PD (2005) Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123**(4)**:** 607-620

Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15**(2)**:** 185-197

Mevel-Ninio M, Pelisson A, Kinder J, Campos AR, Bucheton A (2007) The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. *Genetics* **175**(4)**:** 1615-1624

Miyoshi K, Tsukumo H, Nagami T, Siomi H, Siomi MC (2005) Slicer function of Drosophila Argonautes and its involvement in RISC formation. *Genes Dev* **19**(23)**:** 2837-2848

Mui Chan C, Zhou C, Brunzelle JS, Huang RH (2009) Structural and biochemical insights into 2'-O-methylation at the 3'-terminal nucleotide of RNA by Hen1. *Proc Natl Acad Sci U S A* **106**(42)**:** 17699-17704

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**(Pt 3)**:** 240-255

Nagao A, Mituyama T, Huang H, Chen D, Siomi MC, Siomi H (2010) Biogenesis pathways of piRNAs loaded onto AGO3 in the Drosophila testis. *RNA* **16**(12)**:** 2503-2515

Nishida KM, Okada TN, Kawamura T, Mituyama T, Kawamura Y, Inagaki S, Huang H, Chen D, Kodama T, Siomi H, Siomi MC (2009) Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in Drosophila germlines. *EMBO J* **28**(24)**:** 3820-3831

Nishida KM, Saito K, Mori T, Kawamura Y, Nagami-Okada T, Inagaki S, Siomi H, Siomi MC (2007) Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. *RNA* **13**(11)**:** 1911-1922

Ohara T, Sakaguchi Y, Suzuki T, Ueda H, Miyauchi K (2007) The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. *Nat Struct Mol Biol* **14**(4)**:** 349-350

Okada C, Yamashita E, Lee SJ, Shibata S, Katahira J, Nakagawa A, Yoneda Y, Tsukihara T (2009) A high-resolution structure of the pre-microRNA nuclear export machinery. *Science* **326**(5957)**:** 1275-1279

Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, Lai EC (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* **453**(7196)**:** 803-806

Olivieri D, Sykora MM, Sachidanandam R, Mechtler K, Brennecke J (2010) An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in Drosophila. *EMBO J* **29**(19)**:** 3301-3317

Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* **315**(5809)**:** 241-244

Pardue ML, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL (2005) Two retrotransposons maintain telomeres in Drosophila. *Chromosome Res* **13**(5)**:** 443-453

Parker JS, Barford D (2006) Argonaute: A scaffold for the function of short regulatory RNAs. *Trends Biochem Sci* **31**(11)**:** 622-630

Parker JS, Roe SM, Barford D (2004) Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J* **23**(24)**:** 4727-4737

Parker JS, Roe SM, Barford D (2005) Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature* **434**(7033)**:** 663-666

Pelisson A, Sarot E, Payen-Groschene G, Bucheton A (2007) A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the Drosophila ovary. *J Virol* **81**(4)**:** 1951-1960

Pelisson A, Song SU, Prud'homme N, Smith PA, Bucheton A, Corces VG (1994) Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *EMBO J* **13**(18)**:** 4401-4411

Perrakis A, Harkiolaki M, Wilson KS, Lamzin VS (2001) ARP/wARP and molecular replacement. *Acta Crystallogr D Biol Crystallogr* **57**(Pt 10)**:** 1445-1450

Project CC (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**(Pt 5)**:** 760-763

Prud'homme N, Gans M, Masson M, Terzian C, Bucheton A (1995) Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics* **139**(2)**:** 697-711

Qi H, Watanabe T, Ku HY, Liu N, Zhong M, Lin H (2010) The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. *J Biol Chem* **286**(5)**:** 3789-3797

Rand TA, Petersen S, Du F, Wang X (2005) Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* **123**(4)**:** 621-629

Read RJ (2001) Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr* **57**(Pt 10)**:** 1373-1382

Reinhart BJ, Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**(5588)**:** 1831

Reuter M, Chuma S, Tanaka T, Franz T, Stark A, Pillai RS (2009) Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. *Nat Struct Mol Biol* **16**(6)**:** 639-646

Rivas FV, Tolia NH, Song JJ, Aragon JP, Liu J, Hannon GJ, Joshua-Tor L (2005) Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol* **12**(4)**:** 340-349

Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC (2009) A broadly conserved pathway generates 3'UTR-directed primary piRNAs. *Curr Biol* **19**(24)**:** 2066-2076

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**(6)**:** 1193-1207

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* **17**(12)**:** 1850-1864

Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, Kotani H, Asai K, Siomi H, Siomi MC (2009) A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature* **461**(7268)**:** 1296-1299

Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC (2010) Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. *Genes Dev* **24**(22)**:** 2493-2498

Saito K, Ishizuka A, Siomi H, Siomi MC (2005) Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol* **3**(7)**:** e235

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev* **20**(16)**:** 2214-2222

Saito K, Sakaguchi Y, Suzuki T, Siomi H, Siomi MC (2007) Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes Dev* **21**(13)**:** 1603-1608

Sambrook and Russell (2001) Molecular Cloning: A Laboratory Manual. Third edition, cold spring harbor laboratory press, cold spring harbor, New York

Sarot E, Payen-Groschene G, Bucheton A, Pelisson A (2004) Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. *Genetics* **166**(3)**:** 1313-1321

Schmidt A, Palumbo G, Bozzetti MP, Tritto P, Pimpinelli S, Schafer U (1999) Genetic and molecular characterization of sting, a gene involved in crystal formation and meiotic drive in the male germ line of Drosophila melanogaster. *Genetics* **151**(2)**:** 749-760

Shoji M, Tanaka T, Hosokawa M, Reuter M, Stark A, Kato Y, Kondoh G, Okawa K, Chujo T, Suzuki T, Hata K, Martin SL, Noce T, Kuramochi-Miyagawa S, Nakano T, Sasaki H, Pillai RS, Nakatsuji N, Chuma S (2009) The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline. *Dev Cell* **17**(6)**:** 775-787

Sijen T, Steiner FA, Thijssen KL, Plasterk RH (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* **315**(5809)**:** 244-247

Simon B, Kirkpatrick JP, Eckhardt S, Reuter M, Rocha EA, Andrade-Navarro MA, Sehr P, Pillai RS, Carlomagno T (2011) Recognition of 2'-O-methylated 3'-end of piRNA by the PAZ domain of a Piwi protein. *Structure* **19**(2)**:** 172-180

Song JJ, Liu J, Tolia NH, Schneiderman J, Smith SK, Martienssen RA, Hannon GJ, Joshua-Tor L (2003) The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol* **10**(12)**:** 1026-1032

Song JJ, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305**(5689)**:** 1434-1437

Steiner FA, Okihara KL, Hoogstrate SW, Sijen T, Ketting RF (2009) RDE-1 slicer activity is required only for passenger-strand cleavage during RNAi in Caenorhabditis elegans. *Nat Struct Mol Biol* **16**(2)**:** 207-211

Tian Y, Simanshu DK, Ma JB, Patel DJ (2011) Structural basis for piRNA 2'-O-methylated 3'-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. *Proc Natl Acad Sci U S A* **108**(3)**:** 903-910

Tolia NH, Joshua-Tor L (2007) Slicer and the argonautes. *Nat Chem Biol* **3**(1)**:** 36-43

Vagin VV, Hannon GJ, Aravin AA (2009a) Arginine methylation as a molecular signature of the Piwi small RNA pathway. *Cell Cycle* **8**(24)**:** 4003-4004

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**(5785)**:** 320-324

Vagin VV, Wohlschlegel J, Qu J, Jonsson Z, Huang X, Chuma S, Girard A, Sachidanandam R, Hannon GJ, Aravin AA (2009b) Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev* **23**(15)**:** 1749-1762

Vasileva A, Tiedau D, Firooznia A, Muller-Reichert T, Jessberger R (2009) Tdrd6 is required for spermiogenesis, chromatoid body architecture, and regulation of miRNA expression. *Curr Biol* **19**(8)**:** 630-639

Vilkaitis G, Plotnikova A, Klimasauskas S (2010) Kinetic and functional analysis of the small RNA methyltransferase HEN1: the catalytic domain is essential for preferential modification of duplex RNA. *RNA* **16**(10)**:** 1935-1942

Wang G, Reinke V (2008) A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Curr Biol* **18**(12)**:** 861-867

Wang J, Saxe JP, Tanaka T, Chuma S, Lin H (2009a) Mili interacts with tudor domain-containing protein 1 in regulating spermatogenesis. *Curr Biol* **19**(8)**:** 640-644

Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ (2008) Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* **456**(7224)**:** 921-926

Wang Y, Juranek S, Li H, Sheng G, Wardle GS, Tuschl T, Patel DJ (2009b) Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* **461**(7265)**:** 754-761

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**(7194)**:** 539-543

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**(5)**:** 855-862

Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. *J Appl Cryst* **43:** 186-190

Yan KS, Yan S, Farooq A, Han A, Zeng L, Zhou MM (2003) Structure and conserved RNA binding of the PAZ domain. *Nature* **426**(6965)**:** 468-474

Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* **17**(24)**:** 3011-3016

Yin H, Lin H (2007) An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature* **450**(7167)**:** 304-308

Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW, Steward R, Chen X (2005) Methylation as a crucial step in plant microRNA biogenesis. *Science* **307**(5711)**:** 932-935

Yuan YR, Pei Y, Chen HY, Tuschl T, Patel DJ (2006) A potential protein-RNA recognition event along the RISC-loading pathway from the structure of A. aeolicus Argonaute with externally bound siRNA. *Structure* **14**(10)**:** 1557-1565

Zamore PD, Tuschl T, Sharp PA, Bartel DP (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**(1)**:** 25-33

Zhou R, Czech B, Brennecke J, Sachidanandam R, Wohlschlegel JA, Perrimon N, Hannon GJ (2009) Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. *RNA* **15**(10)**:** 1886-1895

# Appendix

Expression vector details:

1. pTWO-E vector:
    - Minimal His6 tag, 3C protease cleavable
    - Based on pET-17b (Novagen), expression in *E. coli*
    - Unifled multiple cloning site
    - Ampicilin resistance
    - 3.4 kb in length
2. pGEX-6P-1 vector (GE-Healthcare)
3. pET-30a vector (Novagen)
4. pOPINS vector:
    - His-Sumo Tag, Sumo protease cleavable
    - Based on pET28a (Novagen), expression in *E. coli*
    - Unifled multiple cloning site
    - Kanamycin resistance
    - 5 kb in length